Longitudinal Approaches for Metagenomic Characterization of the Puget Sound for Environmental Health Surveillance

Jessica Youngblood

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2013

Committee:

Elaine Faustman (Chair)

Alison Cullen

James Wallace

Program Authorized to Offer Degree:

Environmental and Occupational Health Sciences

University of Washington

Abstract

Longitudinal Approaches for Metagenomic Characterization of the Puget Sound for Environmental Health Surveillance

Jessica Youngblood

Chair of Supervisory Committee:

Dr. Elaine Faustman

Department of Environmental and Occupational Health Sciences

The marine environment is the largest, most diverse and influential ecosystem on Earth. Still largely unexplored, the foundation for further ocean exploration begins with the most abundant and productive life forms in the ocean, the microbial communities. Microbes are essential to all life and play an intimate role in ecosystem function and environmental health. Microbial diversity and community function are important metrics that can be used to monitor and predict environmental changes. Standard lab techniques used for environmental microbial assessment are limited in scope and high-throughput, comprehensive approaches offer a tremendous opportunity to expand our estimates and monitoring of microbial diversity. Metagenomic profiling offers a sensitive approach to evaluate intact community genomes for the novel detection and characterization of microbial populations. Its gene-based, population level surveillance provides advanced

insight into uncultured organisms broadening our understanding of microbial environments, community composition and functional potential. In addition to ecological relevance, metagenomic surveillance creates translational research opportunities for monitoring environmentally hosted human health determinants. The objective of this study was to further define the Puget Sound metagenome by expanding our assessment of coastal areas and their environmental signals of human impact and environmental health relevance. This is the second metagenomic study of the Puget Sound, and includes the addition and characterization of seven metagenomes, comprising a total of 14 samples from 10 different locations including a proximal wastewater treatment plant that discharges effluent into the Puget Sound. This longitudinal study uses 454 next generation sequencing, field metadata, and bioinformatic analysis to profile the surface water bacterial communities of the Puget Sound, both temporally and spatially, to characterize community composition, functional potential, and human health determinants. Our results revealed the high reproducibility and discriminatory capabilities of metagenomic profiling. Repeat samples taken approximately a year apart exhibited highly similar composition, while repeat samples taken during different seasons displayed considerable compositional differences suggesting that environmental conditions influence taxonomic relative abundance. Comparative analysis of all metagenomes exposed significant differences in both microbial diversity and human health determinants across a gradient of anthropogenic impact. These results demonstrate our improved characterization of the Puget Sound Metagenome and build capacity toward future bioinformatic applications in environmental monitoring and public health decision-making.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Microbial Community

Microorganisms are included in Archaea, Bacteria, Eukaryota and Virus domains and are our primary source and recycler of nutrients, neutralization and degradation of toxins [1]. Microbes are essential to all life and their community function is dependent upon their environment, as revealed by their rapid response to environmental changes and intimate role in ecosystem function [2]. Given their universal presence, quick response time, and vast ability to reproduce, microbial diversity indices have become a sensitive measure of change and the health state of an environment [3]. Standard laboratory based culture and isolation techniques used for microbial assessment are limited in scope, and it is estimated that >99% of environmental organisms have not been cultured [4]. Understanding microbial communities is fundamental to the maintenance and conservation of our marine ecosystems and environmental health and high-throughput, comprehensive approaches must continue to be endorsed to appropriately estimate, record, and conserve microbial diversity. A prevalent yet emerging area of microbial assessment, is environmental genomics, also known as Metagenomics.

Metagenomics

Metagenomics is a culture and isolation independent method of exploring genetic material recovered directly from heterogeneous environments; providing access to uncultured microbes and revealing insight into community composition, functional capabilities, evolutionary rates and health of an ecosystem. The study of metagenomes provides potential applications in environmental health monitoring, remediation and generation of beneficial, sustainable products. Using metagenomic applications to study

microbial populations can provide valuable information characterizing our most influential environments, such as our marine ecosystems.

Marine Environment

The marine habitat is the largest ecosystem on earth, it covers approximately 70% of the earth's surface, contains 97% of the planet's water, and plays an integral role in many of the Earth's most influential systems including climate and weather [5]. Its overall impact, influence and need for sustainability cannot be understated, yet due to its vast dimensions, diverse populations and high density more than 95% of the underwater world remains immeasurable and unexplored. Microbes are the most abundant life forms in the ocean, account for approximately 90% of the oceans biomass and are responsible for 98% of the ocean's primary production making them an appropriate foundation for ocean exploration. Of particular microbial interest, are Prokaryotic organisms in the Bacterial domain.

Bacterial Community

Bacteria are distributed ubiquitously throughout nature, account for the majority of the earth's biomass [6] and are a major contributing source of human disease and mortality [7]. Worldwide, bacterial pathogens impose a heavy burden of disease and fear, as treatment options are becoming more and more compromised as a result of anthropogenic impact and bacteria's acquisition, evolution and progression of antibiotic resistance genes. Using metagenomics to explore the marine bacterial diversity can provide a sensitive source for detection, identification and prediction of environmental conditions and change motioning a preventative measure against pathogenic bacteria. By exploring two common diversity measures, species richness and the evenness of species distributions, we can

begin to assess the scope of anthropogenic contamination, a potentially powerful agent of selection acting upon aquatic organisms [8]. Diversity of marine ecosystems is integral to their stability and function and an extensive meta-analysis on marine communities by Johnston et al. found that anthropogenic contamination of marine habitats was frequently associated with reductions in species richness and/or species evenness (increased dominance of tolerant species) with pollutant effects resulting an average reduction in species richness of 30-50% [9]. In order to fully understand the spectrum of effects caused by contaminants, we must first establish baseline populations of marine habitats. Insights from these assessment studies can expose vast amounts of information on indigenous bacterial communities, revealing compositional, functional and environmental preferences that suggest the key species and roles necessary to sustain a functioning environment. Identifying local conditions along a gradient of impact will allow us to further evaluate the causes and effects of adverse pressures and isolate biomarkers capable of assessing human impacts on marine ecosystems and marine impacts on human health.

Coastal Environment

An ecosystem of enormous importance to public health that is responsible for the inhabitance of 163.8 million people, approximately 52 percent of the United States population is the coastal watershed area [10]. These highly accessible areas are among the most diverse and productive on earth [11] and serve a substantial environmental and economic role. Although the coastal area is less than 20 percent of total land area in the United States, its population density far exceeds the entire nations, and is only predicted to keep increasing with estimates of a 9 percent increase by 2020 [10]. As population density continues to grow, the challenges and threats of protecting the coastal environments, as

well as the human population, will become far more demanding [12]. Of significant concern to coastal environments, are the imminent effects of global warming; causing sea levels to rise, ocean acidification, climate change and its immediate consequences resulting in the increasing intensity and frequency of devastating storms. Other activities that are profoundly influencing the coastal waterways include over-harvesting, real-estate development, wastewater, stormwater, agricultural run-off and industrial contamination. Of particular interest, is the coastal water contamination resulting from raw and treated wastewater effluent.

Wastewater

Recently, as a result of our economic advances and steady population growth, the environmental load of domestic, industrial and commercial wastewater has substantially increased [13]. Over the last decade, wastewater composition has increased in both biological and chemical agents including antibiotic resistant genes/bacteria and persistent pharmaceuticals, pesticides and heavy metals [14, 15]. The overall abundance and diversity of these foreign agents found in surrounding waterways are an indication of wastewater impact on the environment. As a result of this biological and chemical pollution, there has been a considerable increase in surface water contamination and ecosystem degradation resulting in the enhancement of potential human health determinants such as antibiotic resistance genes, virulence factors and pathogenic bacteria.

Public Health Implications

As environmental sequencing technologies become more affordable and resourceful, additional opportunities for metagenomics applications develop. These studies over time will disclose vast amounts of information on novel organisms and metabolic

processes allowing scientists to better characterize microbial phenotypes and environmental niches; signifying key species and roles fundamental in sustaining an efficient, functional environment. In addition to its ecological relevance, advanced metagenomic surveillance of environments allows for translational research into human health by increasing the limits of detection and quantification of human health determinates. Using a broad scale screening approach, an Antibiotic Resistance Determinants (ARD) index, as first established by Port et al in "Metagenomics and Antibiotic Resistance surveillance" we can begin to explore new approaches at the population level for environmental health monitoring. The ARD index is a highly sensitive decision tool and is based on the method of using metagenomics in combination with 454-pyrosequencing and bioinformatics analysis to survey antibiotic resistance determinants from intact marine communities and WWTP genomes. The index provides a method of characterizing potential ARDs in the environment and is a key component in the transition of incorporating environmental genomic data into public health decision and management. Natural environments are hospitable reservoirs for antibiotic resistant genes that have emerged and evolved from natural and anthropogenic sources.

Antibiotic Resistance

For centuries, people have benefited from the use of antimicrobials as a treatment to fight infection, and for the past seven decades antibiotics have been successfully synthesized and manufactured serving as the primary method of treatment; greatly reducing illnesses and infectious diseases [16]. When used appropriately, antibiotics serve a vast purpose by killing or inhibiting the growth of harmful microorganisms, ultimately preventing and saving countless lives. Unfortunately, as a result of the widespread use,

overuse and misuse we are now faced with the challenge of bacteria derived resistance, rendering drugs less effective and in some cases useless [17]. Through our extensive efforts at combating microbial life, we have essentially created stronger, thriving life relentless in its survival, boundaries, and persistence; life that is present in all ecosystems, including our marine environment. Currently, assessing environmental health determinants has been restricted to culture and isolation based methods of individual species of historical concern. Given the evolution and progression of these potentially harmful agents, new monitoring tools capable of population assessments are necessary to provide the adequate detection, coverage and prevention.

Antibiotic Resistance Determinants Index

In order to adequately assess the potential levels of Antibiotic Resistance (AR) in the environment at the community level, we must first consider all of its potential determinants. Factors that are included in ARD Index, that potentially play a role in evolution, progression and acquisition of AR are classified into three categories that provide the ecological context and etiology for resistance potential. These categories include: antibiotic resistance genes, gene-transfer potential and pathogenicity potential and are further defined by sub-categories. The antibiotic resistance gene category includes the ARGs and the Metal Resistance Genes (MRGs). The widespread increase and spread of AR is suggested to be driven largely by the selective pressures and gene linkages of antibiotics, heavy metals and disinfectants [13, 16, 18]. Gene-transfer potential contains the mobile genetic elements (MGEs) and is further divided into: plasmids, transposable elements (TEs) and phages. According to the definition provided by Frost et al., MGEs are DNA segments that encode enzymes and proteins that mediate DNA movement within and

between other genomes and are often involved in horizontal gene transfer (HGT) of ARGs [19, 20]. The agents typically known to mediate DNA movement include: plasmids, transposable elements and phages; all agents that can be found in environmental metagenomes. Plasmids, the main vectors for HGT, are stable, self-replicating functional genetic elements [20] and typical carriers of ARGs. Transposable elements (TEs) are mobile segments of DNA that are capable of genomic repositioning, (transferring ARGs) and size alteration and can occupy a high proportion of a species genome volume [21]. Phages (virus of bacteria), are among the most abundant (10^6 -10^9 particles/ml seawater), diverse biological agents found in the marine environment [22] and play a major role in microbial community diversity through HGT and direct bacteria infection [23]. Lastly, the pathogenicity prolife includes the pathogenic bacteria genera classified by the Microbial Rosetta Stone Database [24]. These genera include opportunistic pathogens that are capable of causing serious illness to human populations, in particular immune-compromised and sensitive hosts.

Puget Sound Estuary

This study was the second metagenomic survey of the Puget Sound, WA and is based off of the initial survey by Port et al. "Metagenomic Profiling of Microbial Composition and Antibiotic Resistance Determinants in the Puget Sound" [25]. The Puget Sound estuary is the second largest estuary in the United States and is an inlet of the Pacific Ocean and is connected by the Strait of Juan de Fuca. This estuary is characterized as a fjord system of flooded glacial valleys, containing expansive areas of deep open water to shallow bays and numerous inlets resulting in high seasonal freshwater input from the Olympic and Cascade Mountain watersheds. The Sound covers around 2,500 miles of shoreline and the Puget

Sound Region provides homes to approximately 67% of Washington State's population [26]. These waterways are an integral part of the Washington community and contribute substantially to the economic stability and growth of the Pacific Northwest. Degradation to these environments is detrimental, and one area of great concern is the significant source of wastewater and stormwater pollution resulting from the large amounts of annual rainfall. These contaminated discharges are threatening the coastal waterways and creating a significant health risks for human and aquatic health and in 2012, approximately 154 million gallons of untreated sewage spilled in Seattle's local waterways [27].

Objectives

The objective of this study was to further define the Puget Sound Metagenome by more effectively addressing coastal areas and their environmental signals of human impact and environmental health relevance. This longitudinal study continued to use 454 next generation sequencing, field metadata, and bioinformatic analysis to profile the surface water bacterial communities of the Puget Sound, both temporally and spatially, to characterize community composition, functional potential, and antibiotic resistance determinants across diverse marine environments. This data will further initiate longitudinal monitoring of Puget Sound and may allude to future areas of environmental health concerns for highly impacted areas of the Puget Sound.

MATERIALS AND METHODS

Ethics Statement

Sampling in the Puget Sound area required no specific permits for the collection of water from off the coastal and open sound marine environments. Open sounds cruise sample stations are monitored by the University of Washington Puget Sound Regional

Synthesis Model (PRISM) program. All necessary permits were obtained for the collection of the wastewater treatment plant (WWTP) effluent sample with permission granted by the West Point Treatment Plant (King County, WA), specifically Betsy Cooper (NPDES Administrator, Wastewater Treatment Division, King County Department of Natural Resources and Parks) and Rick Hammond (Chief Process Analyst, West Point Treatment Plant). Field studies did not involve endangered or protected species.

Study Design

This field-based study is the second generation of metagenomic analysis of the Puget Sound Estuary. The framework and design of this study is based on the methods by Port et al., which provided the baseline framework and working index from which our 2012 sampling study continues to develop [25]. The primary variables investigated were the taxonomic composition, function potential and the antibiotic resistance determinants of surface water bacterial communities along the coastal areas of the Puget Sound. Particular attention was focused on our Nearshore marine areas that are greatly affected by anthropogenic and environmental influences such as: sewer, agricultural, and industrial run-off and their contributing role in the evolution, progression and persistence of human health determinants in the Puget Sound.

Sampling Locations

Figure 1 list all 2010-2012 Puget Sound Locations. Below is a list of the seven 2012 sample locations used in this study, followed with a brief description of their position and relevance.

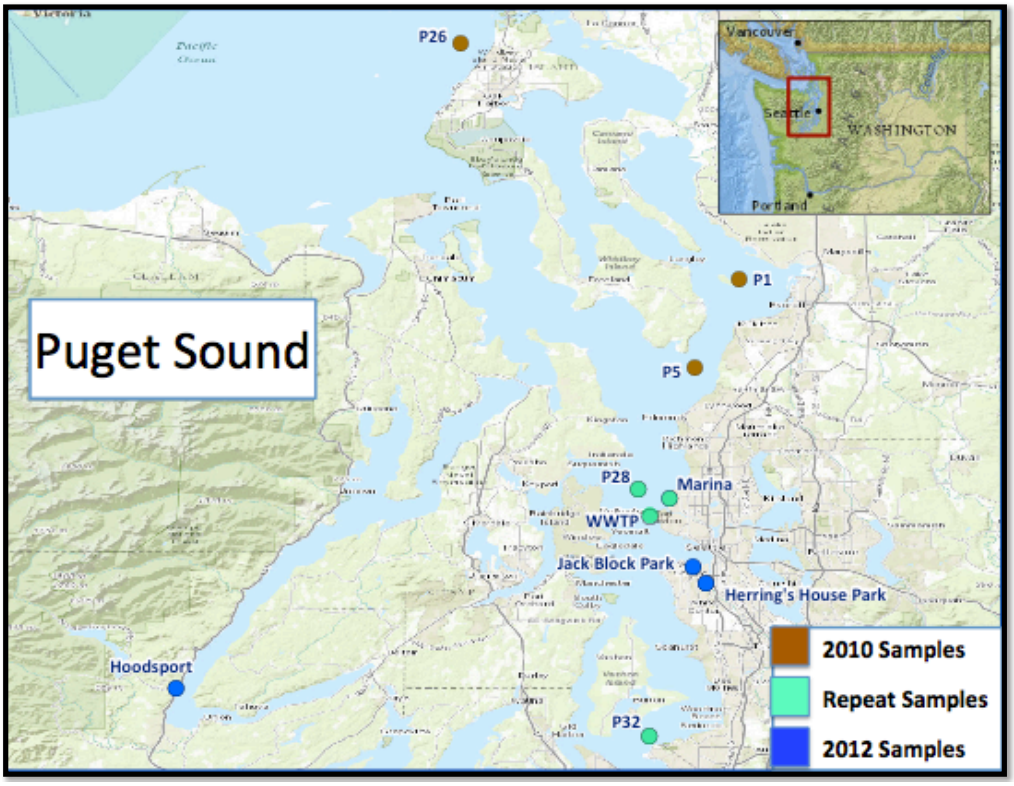

Figure 1. Puget Sound Sampling Map 2010-2012. Refer to Supplementary Table 1 and 2 for the geographic coordinates of 2012 Sampling Sites.

Herring's House Park (HHP)- Lower Duwamish Waterway

A recently restored public waterfront area located on mile 2 of the five-mile stretch of the Lower Duwamish River, an EPA recognized Superfund Site circa 2001. Prior to industrialization, this upland area was originally a marsh waterway of the Duwamish River until 1929, when the area was developed into to the Seaboard Lumber Mill. For approximately 50 years, this area was filled with waste and subjected to the foreign materials contaminated with total petroleum hydrocarbons (TPHs), lead, mercury, and polycyclic aromatic hydrocarbons (PAHs) [28] Fortunately, in 1998 the Intertidal Habitat Restoration Project was initiated and in 2001, after approximately three years of restoration, Herring's House Park opened.

The waterfront area of Herring's House Park is still highly contaminated by industrial waste, runoff, and combined sewer overflows (CSOs) owned and operated by Seattle Public Utilities, King County, and the Port of Seattle. Sewage and stormwater are normally collected in combined sewer basins and flow to a treatment plant however, during large storms, combined sewer overflows drain directly into the Duwamish Waterway.

Jack Block Park (JBP)- Elliot Bay

This public waterfront area resides on the 1994 designated Pacific Sound Resources Superfund Site, a former creosote wood treating facility [29, 30]. After years of remediation, including contaminated waste removal and sediment capping, Jack Block Park, a 13 acre park opened in 1998 as part of the Port of Seattle's redevelopment of Terminal 5 cargo-handling facility. This waterfront area has a close proximity to the mouth of the Duwamish River and central location in Elliot bay. This area exhibits high anthropogenic activity (both residential and commercially), ship trafficking and deposition from the heavily contaminated Duwamish River including the output of over 14 CSO's.

Ingvald J. Gronvold Park (Hoodsport)- Lower Hood Canal

Hoodsport, located on the eastern edge of the Olympic Peninsula, is a densely populated waterfront area on the lower hood canal with a history of chronic exposure to high levels of natural and synthetic organic material (on-site sewage systems, pet and yard waste). Due to the location and topographical constraints (bounded by mountains and steep slopes, underlying impermeable layers), this area receives large volumes of fresh surface waters resulting in increased resonance times (complete turnover takes years), low

mixing, highly stratified areas of temperature and salinity and low levels of dissolved oxygen, (commonly defined as 0.5-3.0 mg/L) [31].

Marina 2012- Shilshole Marina (Repeat Sample)

This area is a repeat location and was chosen based on its public access and it location, just 2 miles north of the WWTP. This public boat ramp is a highly utilized area, with heavy boat traffic and high anthropogenic activity. The marina is located approximately 8 miles north of city of Seattle and includes over 1400 moorage slips (300 liveaboards) and a public plaza.

Station P28 2011 (Repeat Sample)

Collected in 2011 and analyzed in 2012.

Station P32 2011 (Repeat Sample)

Collected in 2011 and analyzed in 2012.

WWTP 2012 Effluent- King County West Point Treatment Plant (Repeat Sample)

The West Point Treatment Plant is part of King County's regional system that treats wastewater for about 1.5 million people and covers 420 square miles in the Puget Sound region treating wastewater from: storm/groundwater (53%), residential (29%), commercial (17%) and industrial (1%) in north King County. The West Point location averages a daily inflow of approximately 133 million gallons with a maximum capacity of 440 million gallons. Wastewater treatments, includes preliminary (removes large debris), primary (sedimentation), secondary treatment (biological consumption of suspended and organic material) and disinfection (chlorination to remove pathogens) leaving water at least eighty-five percent cleaner and destroying the majority of pathogenic bacteria before

water is de-chlorinated and discharged to the Puget Sound via a 3,600 ft. long outfall pipe with an additional 600 ft. diffuser [32].

Sample Collection

As illustrated in Figure 1 and Supplementary Table 1, this study included water samples from six marine environments located throughout the Puget Sound, WA Watershed and one effluent sample from a proximal King County West Point Treatment Plant, a wastewater treatment plant (WWTP) in Seattle, WA. Puget Sound sampling locations for 2012, included three new coastal locations: Lower Duwamish River (Herring's House Park (HHP)), Elliot Bay (Jack Block Park (JBP)) and the Lower Hood Canal (Hoodsport) and four repeat sample locations P28, P32, Marina, and WWTP that were first established by Puget Sound baseline study in 2010 [25]. Open Sound sample locations, stations P28 and P32 were collected aboard the R/V Thomas Thompson on October 10-11, 2011, respectively during a cruise of opportunity with the University of Washington's School of Oceanography. These stations have been monitored since 1998 by the University of Washington Puget Sound Regional Synthesis Model (PRISM) program. The Nearshore samples: JBP, HHP, Hoodsport and the Marina were collected July 9- 12, 2012, respectively during similar tidal episodes (high tide) to maintain consistent sampling strategies between sample locations. The proximal WWTP was collected on August 1st, 2012.

All sample collection and methods were based on the protocol established in the preliminary study by Port et al., and are further described below [25]. It is important when doing longitudinal based studies to keep sampling, methods and analysis as consistent as possible; this includes DNA extraction protocols as metagenomes produced using different DNA protocols may be considered inappropriate for comparative analyses [33].

At each sample location, 80-liters of surface water was collected at approximately 5 meters of the depth. During sampling, physical and chemical parameters were measured at all 2012 sampling locations using Hanna Instruments, HI9829-01/4T multi-parameter water quality meter to measure temperature, salinity, dissolved oxygen, pH, turbidity, total dissolved solids (Supplementary Table 1). In addition, two 100 mL water samples per location were collected to measure *Enterococci* counts, using Enterolert, an Idexx assay [34]. Enterolert, is a microbial indicator assay that detects *Enterococcus*, a type of fecal bacteria, levels in water and is used to identify potential levels of human pathogens in fresh and salt water. This assay is also used by the Washington's Department of Ecology (DOE) BEACH Program [35] to regulate beach closures, and inform the public of potential health risks. Weather (thermal, wind & precipitation), tidal period, surface condition, bottom composition, debris, wildlife counts and visible water quality conditions were observed and recorded. For the cruise samples, physicochemical conditions including temperature, salinity, and dissolved oxygen levels were measured using a conductivity-temperature-depth (CTD) sensor array mounted on a Niskin bottle rosette. Thoroughly assessing the quality and condition of the water during sampling may help to further characterize microbial populations in relation to their environmental conditions through retrospective correlation analysis.

Filtration

Collected surface water from each location was pumped through a Cole-Palmer, U.S.A. peristaltic pump system and sequentially fraction filtered through a 3-µm (147 mm) polycarbonate membrane (Sterlitech, WA) and a 0.2-µm (147 mm) Supor membrane (PALL Corporation, U.S.A.) to obtain the appropriate Bacterial portion. Depending on the

location's water composition, approximately 6 to 40 liters were filtered per set of membranes. After filtration was complete, the remaining bacterial community present on the 0.2 µm Supor membranes was persevered using a sucrose lysis buffer (50 mMTris HCl,40 mMEDTA, and_ 0.75 M_ sucrose), the filters were cut in half and stored at -80°C for downstream DNA extraction.

DNA Extraction and Sequencing

After all samples had been collected, half filters were thawed and cut into fourths. Total metagenomic DNA was isolated and extracted using an enhanced protocol of the PowerWater®DNA Isolation Kit (Mo Bio Laboratories, CA), a kit that is specifically designed for isolation of high quality genomic DNA from filtered environmental water samples removing potentially interfering organic matter. Protocol modifications included: a 10 min water incubation at 65°C after Solution P1 was added to the PowerWater® Bead Tube (alternate lysis method), and prior to the first centrifuge a 110µl of Lysozyme digestion (100mg/ml) was added followed by a water incubation at 37°C for 45 minutes then the additions of 4-µl of RNase digestion (100mg/ml) and water incubation at 37°C for 15 minutes. DNA concentrations (quantity) ranged from 81 to 157 ng/µl per sample and were estimated 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) (Supplementary Table2) and gel electrophoresis showed high molecular weight DNA fragments. Total community DNA was recovered from each sample and extracted DNA was processed using the Roche/454 GS FLX System Titanium platform in the Department of Microbiology at the University of Washington lab. In this study, dead cells and free associated DNA were not removed prior to sequencing, so it is important to

note that some sequences identified may not be associated with live or active cells and further studies using metatranscriptomics or DNA-binding may be advantageous.

Bioinformatic Analysis

Sequenced-based metagenomic studies require many stages of up and down stream quality control to ensure high quality reads essential for accurate community representation and comparative results. Post sequencing, all 454 sequence reads were trimmed to remove barcodes, secondary adapter sequences and reads that contained more than 5% of any one nucleotide. For each sample, all raw sequence were assembled into contiguous, overlapping sequence reads (contigs) with Newbler software v. 2.5.3 (Roche Diagnostics-454 Life Sciences) using a 95% minimum overlap identity and default settings.

Taxonomic and functional annotation for all 454 unassembled reads were based on the internal information provided by the Metagenomics Rapid Annotations based on Subsystems technology (MG-RAST) server version 3.3.3.3 [36]. MG-RAST is an automated analysis platform for metagenomes, and produces taxonomic and functional assignments of sequences by similarity based comparisons against both protein and nucleotide databases. Optional quality-control filters, an internal tool in the MG-RAST pipeline, was applied to the raw sequence data (FASTA format) during submission and prior to annotation to remove duplicate and low quality reads. As established in our preliminary study, reads meeting any of the following criteria were omitted: read length >2 standard deviations from the mean sample read length, reads containing >5 ambiguous basepairs, and all but a single representative of clusters of reads whose first 50 base pairs are identical (de-replication). The MG-RAST v3.3.3 pipeline downstream analysis of the quality-filtered, unassembled reads included taxonomic classification determined by the lowest common ancestor (LCA)

analysis and functional annotations against the SEED database applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 60% identity, and a minimum alignment length of 50 amino acids (Figure 2). The LCA analysis is a conservative approach that reports the lowest common ancestor for the given set of taxonomic annotations from the M5NR database using the NCBP taxonomy tree. The M5NR, the M5 non-redundant protein database is an integration of several databases (GO: Gene Ontology, Greengenes: 16S rRNA Gene Database, JGI: Joint Genome Institute, KEGG: Kyoto Encyclopedia of Genes and Genome, NCBI: National Center for Biotechnology Information, RDP: Ribosomal Database Project, SEED: The SEED Project, SILVA: SILVA rRNA Database Project, UniProt: UniProt Knowledgebase, VBI: Virginia Bioinformatics Institute, and eggNOG: evolutionary genealogy of genes: Non-supervised Orthologous Groups) into one highly annotated, searchable database. The M5NR identifies potential genes by using the BLAT platform (faster alternative to BLAST) to perform similarity searches against its combined database of known proteins. The M5NR part of the Genomic Standards Consortium (GSC) [37] an initiative working towards harmonized characterizations of genomes beneficial for large scale comparative genomic analyses. The SEED classification assigns genes to functional roles, and functional roles are grouped in a tri- level hierarchy of subsystems, with catabolic/anabolic functions at the highest level and specific gene pathways at the functional level [38].

In addition to using the MG-RAST pipeline for taxonomic annotation, we preformed a complimentary, phylogenetic analysis using reads that matched the highly conserved 16S rRNA gene to determine closest related taxa (Figure 2). For this 16S rRNA analysis, an *Escherichia coli* reference sequence was queried against the MG-RAST quality-controlled

unassembled sequence reads using BLASTN (default settings) and subsequent aligned reads were classified by the Ribosomal Database Project (RDP) Classifier (RDP Classifier Version 2.5 trained on 16S rRNA training set 9) using a minimum confidence threshold of 60 % [39]. The RDP Classifier uses 16S rRNA sequences, to provide hierarchal taxonomic annotation from domain to genus based on the a naïve Bayesian rRNA classifier.

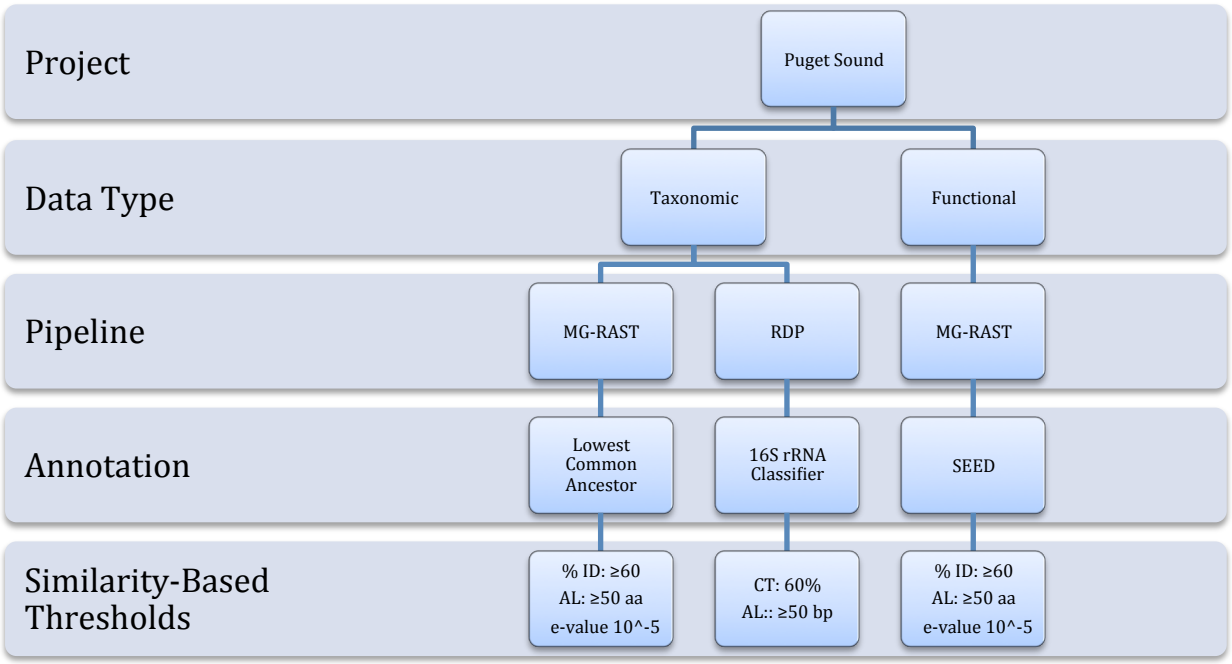

Figure 2. Bioinformatic Analysis for Puget Sound Metagenome Characterization.

Species diversity was annotated using the M5NR database best hit classification, through the MG-RAST pipeline [36] applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 60% identity, and a minimum alignment length of 50 amino acids. Species diversity was then calculated using Shannon diversity index (*H*), a recognized diversity index appropriate to metagenomic studies [40]. The proportion of species relative to the total number of species was calculated, and then multiplied by the natural logarithm of this proportion. The resulting product was totaled across species, and

multiplied by -1. Species evenness was calculated by using Shannon's equitability (E_H), calculated by Shannon diversity index (H) by the natural logarithm of the total number of species. Indices were then converted to true diversity (effective number of species) for comparable interpretation when comparing diversities of different communities by taking the equivalent number, the exponent of the Shannon diversity index (exp(H)) [41].

Methods used to identify components used in the Antibiotic Resistance Determinants Index (ARD Index), were established in Port et al, "Metagenomics and Antibiotic Resistance surveillance" and are briefly explained below (submitted 2013). For consistency, all ARD Index annotations were performed on raw sequence reads that passed the initial MG-RAST QC pipeline. Antibiotic resistance genes (ARGs) were identified using the best-hit approach against an expanded version of the Antibiotic Resistance Gene Database (ARDB). The ARDB+ includes ARDB sequences to a total of 16,085 non-redundant sequences. Potential protein coding sequences were predicted for all sequences using the default setting of MetaGeneMark [42] and predicted proteins were then BLASTP searched against the ARDB+ applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids. ARG sequences were normalized to the total number of sequence reads per sample. Metal resistance genes (MRGs) were identified through the functional annotations of the SEED subsystems hierarchal classification system via MG-RAST pipeline [38]. All genes coding for metal resistance in subsystem Level 3, a subset of the Level 2 "Resistance to antibiotics and toxic compounds" included: "Arsenic Resistance", "Cadmium Resistance", "Cobalt-Zinc-Cadmium Resistance" and "Mercury Resistance Operon" and "Zinc Resistance". Metagenomic reads matching a metal resistance gene with an e-value threshold of 1.0 e-5, a

minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids were retained. Metal resistance gene sequences were normalized to the total subsystem Level 1 count per sample. Plasmid sequences were annotated by searching all sequence reads against the NCBI RefSeq plasmid database (including 13,314 plasmids) using BLASTN [43] applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 95% identity, and a minimum alignment length of 50 base pairs. Plasmid counts were normalized to the total number of sequence reads per sample. Transposable elements (TEs) were identified by comparing all unassembled sequence reads against unique genes annotated as TEs through GenBank (including 430,841 TEs) using BLAST applying a minimum sequence similarity criteria of 80%, and a minimum alignment length of 50 amino acids. TE counts were normalized to the total number of sequenced reads per sample. Phages were taxonomically annotated through the MG-RAST server using the internal LCA pipeline. All unassembled sequence reads matching to the domain Virus applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids. Total phage counts were normalized to the total number of annotated sequences at the domain level per sample. Pathogens were taxonomically identified by comparing the genus level annotations from the Microbial Rosetta Stone Database [24] to the genus level based annotations using the LCA and RDP Classifier based methods. Genus level LCA annotations from the MG-RAST pipeline [36] used a similarity based criteria consisting of an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 95% identity, and a minimum alignment length of 50 amino acids.

Statistical Analysis

Prior to quantification, respective level abundance counts were normalized against total abundance counts (relative abundance) of annotated reads within a sample to account for differences in number of sequences across samples. Effect level distribution of taxonomic and functional composition was produced using the Statistical Analysis of Metagenomic Profiles (STAMP) software package [44]. Only relative abundances 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 and p-values were <0.05 for all comparisons. The Benjamini-Hochberg FDR method was used to correct for the false discovery rate [25]. All taxa were normalized (relative abundance) prior to performing pairwise correlations using Pearson's correlation coefficient.

Metagenome Sequence Accession

Sequence data for the Puget Sound 2010 dataset is available through the 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 and the 2012 dataset will be available under the MG-RAST identification numbers: 4517540.3, 4517541.3, 4517542.3, 4517543.3, 4517544.3, 4517545.3 and 4517546.3.

RESULTS AND DISSCUSSION

In this study, the bacterial fraction (0.2 um- 3.0 um) DNA was isolated from six Puget Sound marine water samples and one proximal WWTP effluent to characterize community composition, functional potential, and probable human health determinants.

2012 Sample Characterization

Overall, 454 metagenome libraries generated 1.45 million sequence reads comprising a total of 539 million base pairs across seven sample locations (Table 1). Prior to preforming sequence annotation, the MG-RAST QC pipeline removed 296072 sequences (~20%) of reads from further analysis, 64% of those sequences (190,014) were classified as clusters of artificially replicated sequences, an occurrence thought to be a byproduct of pyrosequencing [45]. Artificial duplicates from pyrosequencing reads may lead to incorrect interpretation of the abundance of species (overabundance if sequences are retained) but disregarding natural duplicates may also cause an underestimation [46]. In total, high-throughput sequencing performed on surface coastal waters and WWTP effluent DNA generated a total of 1,157,244 quality-filtered sequence reads (5.08 hundred million base pairs) with a mean average read length of 437 (+/- 88) bp that were used for downstream analysis (Table 1). In total, 23,780 contigs (1.6% of total sequences) were assembled and approximately 59% of all assembled contigs were greater than 500 bp in length with the average N50 contig size of 948 bp (samples ranging from 682-1390bp). Bacteria accounted for 97% of all contigs and was predominantly represented by the phyla Proteobacteria (66%), Bacteroidetes (32%) and Actinobacteria (2%).

	Open Sound			Nearshore			WWTP
2012 Seqence Data	**P28 2011**	**P32 2011**	**HHP**	**Hoodsport**	**JBP**	**Marina 2012**	**WWTP 2012**
Number of Sequences	151,376	151,619	419,076	191,398	139,488	246,287	154,072
Artificial Duplicate Reads Sequence Count	20,225	19,196	55,659	24,391	17,661	32,259	20,623
Number of Sequences post MG-RAST Quality Control	119,981	121,716	332,587	152,887	111,775	196,384	121,914
Total Length of Sequence (bp)	52,350,652	52,177,912	148,556,187	65,982,265	48,373,262	86,434,916	54,211,710
Average Read Length (bp)	436 ± 91	428 ± 94	446 ± 82	431 ± 92	432 ± 91	440 ± 86	444 ± 80
Mean GC percent	44 ± 10	43 ± 10	45 ± 9	46 ± 9	45 ± 10	44 ± 9	52 ± 12
N50 Contig	701	682	1390	940	866	1260	797

Table 1: Sequence Summary Statistics for 2012 Samples.

Rarefaction Curves

Rarefaction curves, also referred to as species accumulation curves are estimates of sampling completeness [40]. They provide an estimate of the species diversity based on the sampling depth and determine if the random fraction sample is an adequate representation of the whole community. Our rarefaction curve analysis were produced through the MG-RAST pipeline [36] using their default settings and revealed that although our samples were not exhaustively sampled, they were a satisfactory representation of annotated species in each community (Supplementary Figure 1)

Taxonomic Classification

Approximately, 58% of the 1.16 million quality-controlled (qc), unassembled sequence reads were assigned to protein sequences at the domain level using the MG-RAST pipeline [36] and LCA annotation applying a similarity based criteria consisting of an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 60% identity, and a minimum alignment length of 50 amino acids (Figure 3).

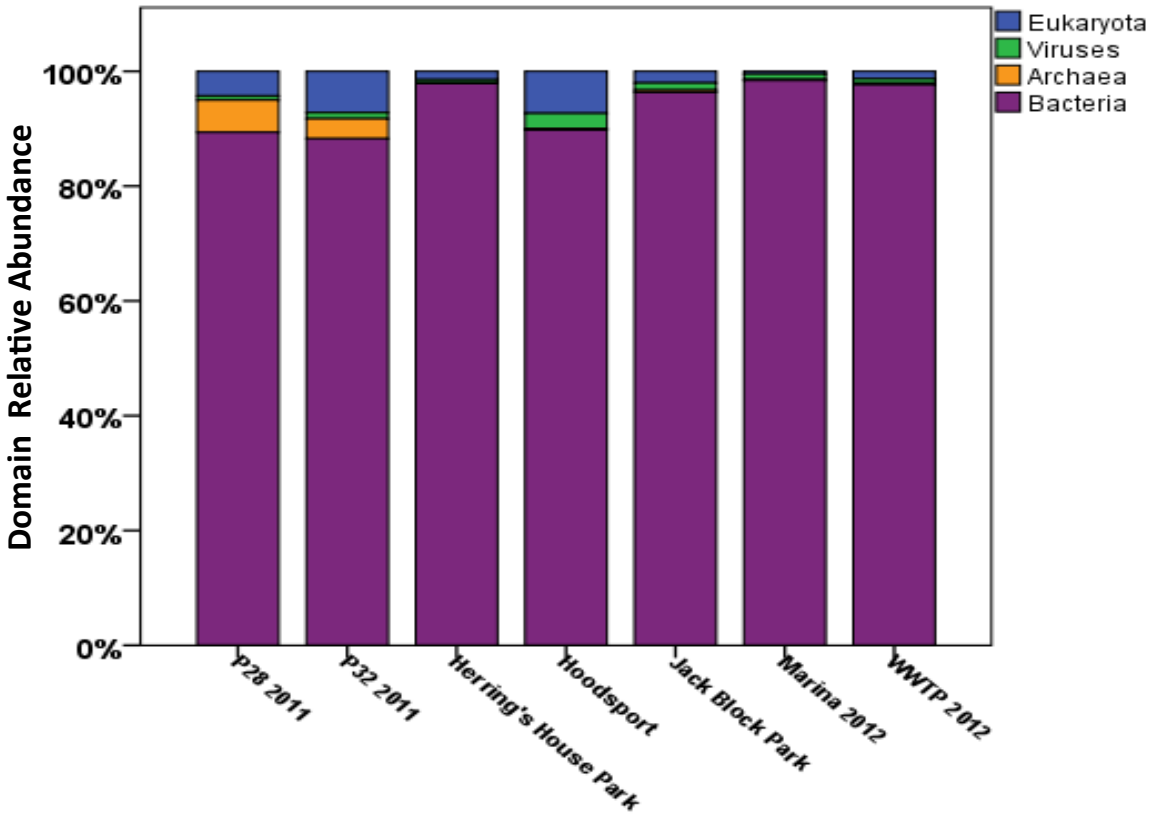

Figure 3. Domain Distribution of 2012 Samples with LCA annotation using the MG-RAST Pipeline. Domain abundance counts were normalized to the total number of Domain abundance counts per sample.

Of the annotated sequence reads, bacteria accounted for 94% (samples ranging from 88-98%) with the remaining portion comprised of Archaea 1.5%, Eukaryota 3.4% and viruses 1.1%. It should be noted, although we preferentially filtered for the bacterial fraction expecting it's domain to be the prevailing annotation, bacterial genes dominate these annotation databases therefore, annotation biases against other domains may result in less representation and adequate distribution. Using the same criteria specified above, approximately 50% of qc unassembled sequence reads were assigned to the phylum level using the conservative MG-RAST LCA annotation (Figure 4).

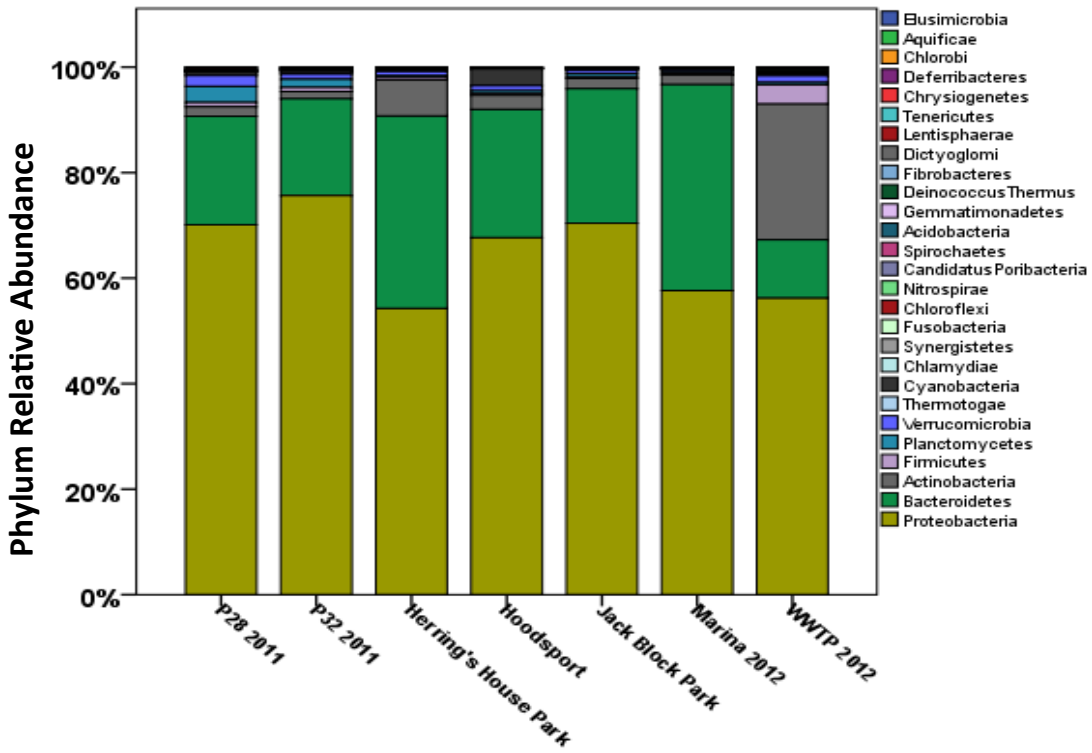

Figure 4. Phylum Distribution of 2012 Samples with LCA annotation using the MG-RAST Pipeline. Phylum abundance counts were normalized to the total number of Phylum abundance counts per sample.

All Puget Sounds communities were predominantly composed of the phyla Proteobacteria, Bacteroidetes, Actinobacteria, Planctomycetes, Verrucomicrobia, Cyanobacteria, and Firmicutes (in ranking order of mean relative abundance). The WWTP 2012 Effluent was composed of Proteobacteria (56%), Actinobacteria (26%), Bacteroidetes (11%), Firmicutes (4%), Verrucomicrobia (1%), Planctomycetes (0.4%), and Cyanobacteria (0.2%). The most abundant phylum, Proteobacteria (~66% mean relative abundance) ranged between 54-76% across all Puget Sound samples, with lowest representation in Herring's House Park. Proteobacteria are found in every environment, they are the most studied phylum and include the majority of recognized agricultural, industrial and medical relevant organisms [47] The Bacteroidetes phylum was the second

leading phyla in all Puget Sound samples, with highest representation in the Nearshore samples (Hoodsport, Jack Block Park, Herring's House Park, and the Marina 2012) ranging from 24-39% relative abundance and had the lowest representation in the WWTP 2012 effluent (11%). Actinobacteria, the third ranked largest phyla across Puget Sound samples, (second in the WWTP effluent), had the highest abundance in the WWTP effluent (26%) and ranged from 1.4- 6.8% in the Puget Sound samples with the lowest representation in Open Sound samples and highest in Herring's House Park. Planctomycetes and Verrucomicrobia both had a Puget Sound mean relative abundance of 1% and were driven by station P28 with respective relative abundances of 3% and 2%, respectively. Hoodsport, an area known for its natural and synthetic induced nutrient rich waters, had the highest representation of Cyanobacteria (Puget Sound mean relative abundance of 0.8%)with a relative abundance of 3.1%. Cyanobacteria (blue-green algae) are typically found in in calm nutrient-rich waters and are known for their explosive reproductive ability to create algal blooms. Some Cyanobacteria species can produce toxins (cyanotoxins) resulting in the production of harmful algal blooms that that are capable of causing serious health effects for animals and humans.

At the class level annotation, the phyla Proteobacteria, was primarily composed of Alphaproteobactera, followed by Gammaproteobacteria, and Betaproteobacteria, a typical distribution found in marine waters. Overall, Alphaproteobacteria accounted for 49% of the mean relative abundance of annotated reads across the Puget Sound samples, and was the dominant class in every sample except for the WWTP 2012 effluent, which was most represented by the class Actinobacteria (29%)(Supplementary Figure 2). Alphaproteobacteria ranged the most between the geographically closest related samples

ranging from 36% in Herring's House Park to 59% in Jack Block Park. The second most abundant class in the Puget Sounds was Flavobacteria, with a mean relative abundance of 21%, (ranging from around 5% relative abundance in WWTP effluent to 40% in the Marina 2012 sample). The next most represented class Gammaproteobacteria, which accounted for around 15% of the total mean relative abundance of all Puget Sound reads, had most representation in the 0pen Sound samples P28 and P32, with a class relative abundance of 20% and 25%, respectively followed by Betaproteobacteria and Actinobacteria (class), at a Puget Sound mean relative abundance of 5% and 3%, respectively. After the class Actinobacteria, the WWTP 2012 effluent was characterized by the Betaproteobacteria (23%), Gammaproteobacteria (16%), and Alphaproteobacteria (12%).

At the order level annotations, Rhodobacterales was dominant in six of the samples, not including the WWTP 2012 effluent (Supplementary Figure 3). Surprisingly, the largest distribution difference was again between the two closest geographically related sampling areas, with a relative abundance of 33% at Herring's House Park and 60% at Jack Block Park. The WWTP effluent, which was most represented by the order Actinomycetales had a relative abundance around 31%. Other highly annotated orders across samples included Flavobacteriales, Burkholderiales and Rickettsiales.

16S rRNA Gene Classification

Although our study was a whole genome survey, a 16S ribosomal RNA (16S rRNA) analyses was performed on all identifiable 16S rRNA genes to determine closest related taxonomy using the Ribosomal Database Project (RDP) Classifier applying a minimum confidence threshold of 60%, and minimum sequence length of 50 base pairs [39]. The 16S rRNA gene is a component of the 30S, small subunit of the prokaryotic ribosomes. Its gene

sequences are useful in phylogenetic studies as they contain hypervariable regions that are highly conserved across Bacteria providing species-specific signatures [48, 49]. The RDP Classifier uses 16S rRNA sequences, to provide hierarchal taxonomic annotation from domain to genus based on the naïve Bayesian rRNA classifier. Overall, 1031 sequences were uploaded to the RDP Classifier, and phylogenetic analysis revealed that approximately 96% of those sequences were annotated at the phylum level, 92% at the class level and 38% at the genus level. Bacterial phylogenetic distributions based on the 16S genes were generally consistent with the LCA annotations using a rank order comparison. As LCA results indicated, the RDP annotated communities were dominated by the phylum Proteobacteria, at a 65% mean relative abundance across Puget Sound samples, with the lowest relative abundance observed in Herring's House Park (57%) and the highest in P32 2011 (77%) (Supplementary Figure 4). Proteobacteria was followed by phylum Bacteroidetes with a 25% mean relative abundance across Puget Sound samples. The WWTP 2012 effluent was also mostly composed of Proteobacteria at 59% relative abundance, but again as we saw using the LCA annotation, Actinobacteria was the second most abundant at a relative abundance of 16%. Class level annotations (Supplementary Figure 5) across the Puget Sound samples were also highly consistent with LCA output, with the most predominant being Alphaproteobacteria (mean relative abundance 32%), in all samples except the WWTP, which was largely composed of the class Betaproteobacteria (26%), Gammaproteobacteria (24%), Actinobacteria (17%); a variation in rank order abundance from earlier results. For the Puget Sounds samples, following Alphaproteobacteria came Flavobacteria, Gammaproteobacteria, with mean relative abundances of 24% and 22%, respectively.

In order to more effectively evaluate taxonomic environment specificity, which is thought to be more pronounced starting at the genus level, and as demonstrated by the large overlap in the higher taxa, (including the phylum, class and order levels), we examined the genus level annotation from the 16S rRNA gene (Figure 5).

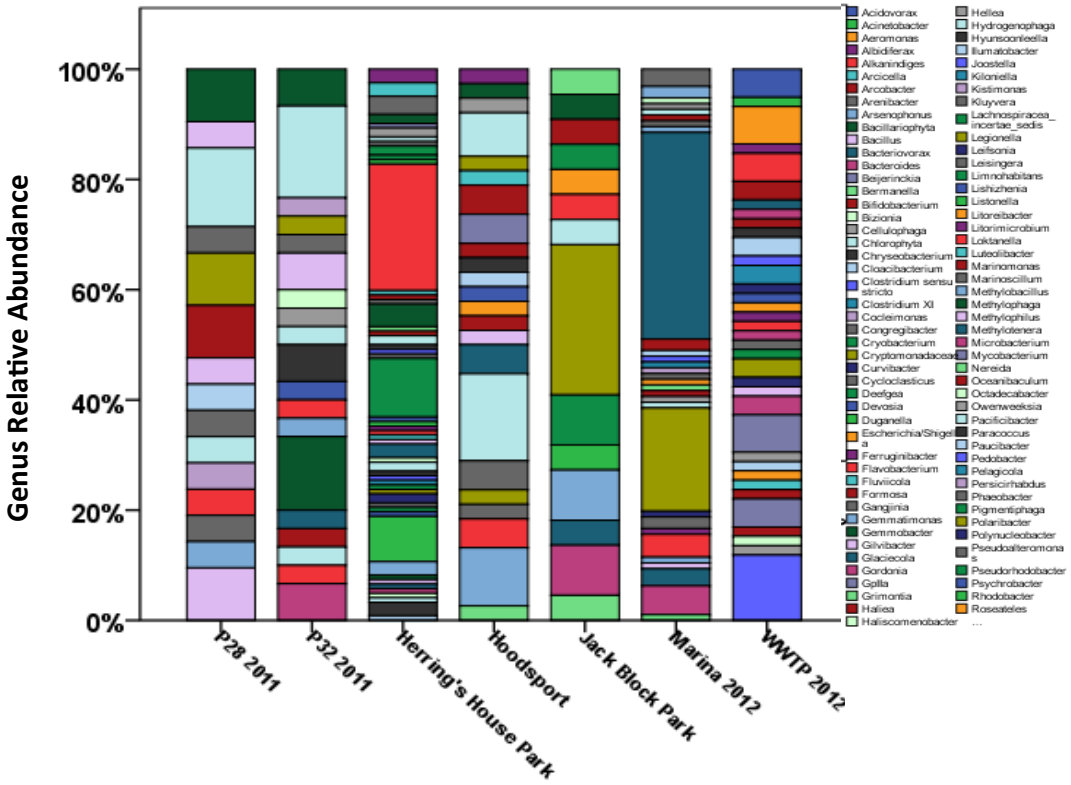

Figure 5. Genus Distribution of 2012 Samples using the 16S rRNA Gene annotations from the Ribosomal Database Project Classifier. Genus abundance counts were normalized to the total number of Genus abundance counts per sample.

Approximately, 338 sequences around 0.03% of total quality-filtered sequences were annotated down to the genus level using the RDP Classifier. Each individual sample harbored a range of genera, with Jack Block Park, represented by 14 genera, and Herring's House Park with 51 genera. Herring's House Park was dominated by *Flavobacterium*, (from the phylum Bacteroidetes), accounting for 23% of genera diversity. *Flavobacterium*, is a genus know for its opportunistic pathogens and mining down to the species level

annotation using the M5NR database through MG-RAST, we found that the species *Flavobacterium psychrophilum (*causative agent in Bacterial Cold Water disease) was accountable for 33% of species diversity of *Flavobacterium* [50]. Bacterial Cold Water Disease (BCWD) is a severe fish disease and is responsible for high morbidity and mortality rates in the Salmon population. The WWTP 2012 effluent, represented by 36 genera, had the most heterogeneous mixed genus sample, with the top genera *Zoogloea* (12%) and Mycobacterium (7%). The Marina 2012 was composed of 30 different genera and was most abundant in *Glaciecola* 38, and *Polaribacter* 19%, which was also most dominant in Jack Block Park 27%. Our Hoodsport sample was dominated by the genus *Pacificibacter*, 16%, and included a total of 23 genera.

Functional Composition

Approximately, 45% of the 1.16 million quality-controlled (qc), unassembled sequence reads were annotated with assigned function using SEED subsystems hierarchal classification levels through the MG-RAST pipeline using an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 60% identity, and a minimum alignment length of 50 amino acids [36, 38]. The SEED classification assigns genes to functional roles, and functional roles are grouped in a tri- level hierarchy of subsystems. SEED annotations resulted in the following subsystem distributions: Level 1 (28 categories), Level 2 (459 categories), Level 3 (1134 categories), and functional roles containing 11,036 functions as of May 2013 (numbers are a reflection of searching against the Puget Sound 2012 dataset and may not comprise all categories and or functions). It should also be noted, a functional role can occur in different subsystems, and therefore a function may be represented more than once in a sample. Using subsystems level 1 classification, potential metabolic

pathways among all Puget Sound samples and the WWTP 2012 effluent included highly similar distributions across all 28 categories (Figure 6).

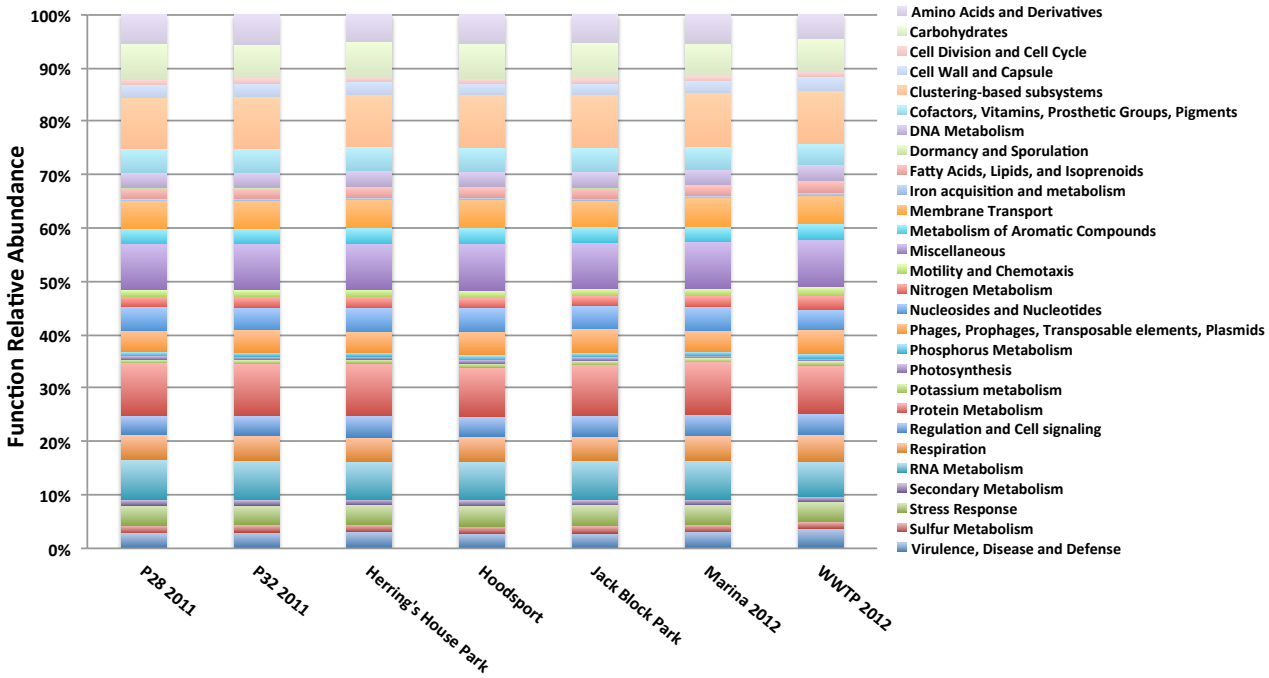

Figure 6. Functional Distribution of 2012 Samples with SEED subsystems Level 1 annotation using the MG-RAST Pipeline. Functional abundance counts were normalized to the total number of Level 1 abundance counts per sample.

Overall, these results suggest that there is high level of functional diversity among all samples and given that we see similar distribution in relative abundance across all samples, it appears that the higher (more general) levels of classification are less sample specific. Also, since only half of our data could be functionally annotated, we may be missing a large component of community function that could play a role in determining significant differences in sample distributions. Due to the current limitations in using similarity-based annotations, novel and poorly referenced or characterized functions that play significant roles may remain overlooked or misclassified and ultimately lower our overall ability to determine predominant functional capabilities. Given the number of

categories and similar distribution at subsystems Level 1 we decided that for the scope of this project, our investigation would focus on categories with potential Public Health relevance beginning in subsystems Level 1 with: “Phages, Prophages, Transposable elements, Plasmids”, “Stress Response” and “Virulence, Disease and Defense”, in order of overall mean relative abundance (12, 14 and 15 respectively out of the 28 Level 1 categories). Overall, these three categories accounted for 10-12% of each sample’s subsystems Level 1 functional annotations (Supplementary Figure 6). Overall, the WWTP 2012 effluent had a slightly larger lead in overall abundance (12%), and open sounds station P28 2011 was least abundant at just under 11%. Further investigation into subsystems level 2, determined that the main drivers behind the Level 1 distributions include categories “Phages, Prophages” (“Phages, Prophages, Transposable elements, Plasmids”), “Oxidative Stress” (“Stress Response”), and “Resistance to antibiotics and toxic compounds” (Virulence, Disease and Defense) (Supplementary Figure 7).

Using these three subsystem Level 3 categories, we determined representative taxonomic affiliations (Figure 7) using the LCA annotation through MG-RAST [36] applying a similarity based criteria consisting of an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 60% identity, and a minimum alignment length of 50 amino acids. All annotated taxa abundances were normalized to total abundance per sample.

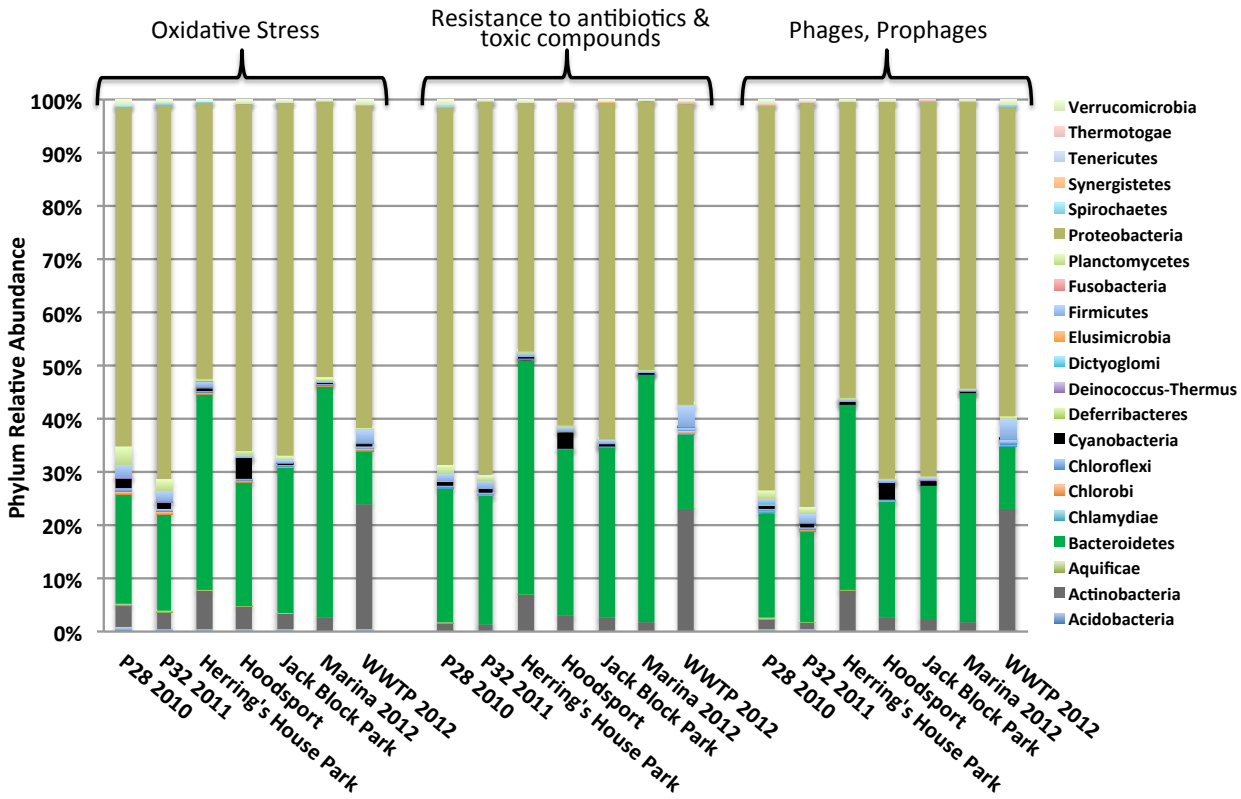

Figure 7. Taxonomic Affiliation with Function of 2012 Sampling Locations. Taxonomic affiliations with Public Health relevant SEED subsystems Level 2 function from LCA annotation using the MG-RAST Pipeline.

Across all three categories, the Bacteria domain was dominant and mainly represented by the phyla Proteobacteria, Bacteroidetes, Actinobacteria, Verrucomicrobia, Cyanobacteria and Firmicutes. Proteobacteria accounted for at least 45% of all the annotated reads in each of the Level 2 categories: "Phages, Prophages", "Oxidative Stress", and "Resistance to antibiotics and toxic compounds". Actinobacteria was more abundant in samples containing lower salinity levels, and was highest in the WWTP, followed by Herring's House Park (HHP). The phylum Cyanobacteria was manly associated with the Hoodsport sample, and was the among the top four most abundant phyla in each of the three seed categories. At the class level annotation, the main driver for Proteobacteria across all three categories was Alphaproteobacteria in the Puget Sound samples followed

by Gammaproteobacteria and Betaproteobacteria in all samples except for HHP where Betaproteobacteria was more abundant than Gammaproteobacteria. In the WWTP, Proteobacteria was primarily composed of Betaproteobacteria, Gammaproteobacteria and then Alphaproteobacteria, (opposite of the majority of Puget Sound samples). In all 2012 samples, the phylum Bacteroidetes was represented by Flavobacteria and Actinobacteria by the class Actinobacteria.

Environmental Metadata

One of the objectives of this study was to further investigate the significant and biologically relevant correlations between annotated metagenomic data and marine associated metadata. Advanced characterization of environments can help define the physical, chemical and biological parameters that play a contributing role in the taxonomic and functional composition of microbial communities. The significance of metadata has become increasingly important and many metagenomic databases are now requiring this valuable information in order to make data publicly available and formatted for optimal comparisons.

In the Puget Sound 2012 sampling, metadata was measured at five locations to assess environmental conditions and water quality using the commonly measured parameters pH, temperature, dissolved oxygen (DO), salinity, turbidity and Enterococcus levels (Supplementary Table 1). In addition to measuring water parameters, weather (thermal, wind & precipitation) and site conditions (tide, wave, bottom composition, debris, wildlife & water quality conditions) were observed. Lastly, in order to characterize the potential public health risk of the 2012 locations (using culture-based methods) we used a standard microbial indicator assay to detect *Enterococci* contamination. *Enterococci*

levels indicate the potential human health risk in recreational waters and they are the preferred fecal indicator bacteria in marine environments due to their increased ability to survive in salt water. This method is supported by the EPA's Recreational Water Quality Standards (RWQS), and has been successful in reducing human exposure to potentially harmful pathogens. These standards require routine water testing and recently released criteria recommendations of the EPA's RWQS (Fall 2012) suggest that coastal waters designated for primary contact recreation use should not exceed: Recommendation 1: a geometric mean (GM) of 35 colony forming units (CFU) per 100 mL (estimated illness rate 36/1,000) or Recommendation 2: a GM of 30 CFU per 100 mL (estimated illness rate 32/1,000) in any 30-day interval [51]. In comparison for secondary contact recreation water, the Washington State Department of Ecology *Enterococci* organism levels must not exceed a GM value of 70 CFUs per 100 mL, with not more than 10 percent of all samples or any single sample to exceed the GM value of 208 CFUs per 100 mL [52].

Water quality assessment is an important environmental monitoring tool used for both ecological and environmental health assessment and thoroughly assessing the quality and condition of the water during sampling increases the characterization of community profiles and opportunities or marine biomarker development. According to the "Water Quality Standards for Surface Waters of the State of Washington", all water quality parameters (including pH, temperature, dissolved oxygen, salinity, turbidity and bacteria) for Puget Sound samples were within the Extraordinary to fair category range however, the WWTP 2012 effluent had exceedingly high levels of the *Enterococci*. Although *Enterococci* organism levels were within the criteria recommendations for coastal locations, we still observed higher amounts in our Herring's House Park and Marina 2012 sample, potentially

indicating environments of increased anthropogenic impact. Due to the limited number of environmental samples with metadata, further analysis of the statistically significant correlations made between the taxonomic composition of predominant taxa and metadata is needed.

Comparative Analyses of All Samples

The objective of this study was to further define the Puget Sound metagenome by continuing to monitor and build upon the taxonomic composition, functional potential and comparative analysis of the surface water bacterial communities along the coastal areas of the Puget Sound and to more effectively address the potential environmental signals of human and biological impact that may be relevant to human health. Thus far we have characterized our latest 2012 samples, determining microbial composition, functional diversity and potential taxa of public health concern. Using our entire metagenomic sample collection of 14 samples, including 12 samples from the Puget Sound watershed and two samples from a WWTP effluent, we further evaluated the Puget Sound Metagenome signature. In order to explore if a gradient of human impact exist between the Open Sound, Nearshore and WWTP effluent samples, we taxonomically and functionally compared all Puget Sounds Samples (including the WWTP effluents) relative to a two sample averaged Open Ocean metagenomic sample, the Sargasso Sea (GSOO1b, GSOO1d) from the Global Ocean Sampling (GOS) Expedition [53]. The GOS Expedition is an ocean exploration genome project developed to characterize microbial diversity in marine communities around the world. Using an Open Ocean sample allows us to investigate the anthropogenic impact on our local waterways compared to a more isolated marine environment, the Sargasso Sea. Identifying local conditions along a gradient of impact will allow us to further

evaluate the causes and effects of adverse pressures and isolate biomarkers capable of assessing human impacts on marine ecosystems and marine impacts on human health. Only taxa and functions with a difference of proportions (percent difference in relative abundance between relative sample and Sargasso Sea) greater than 1% were shown. Positive values indicate higher relative abundance in Puget Sound and WWTP samples, and negative values indicate higher relative abundance in the Sargasso Sea (P-values were < 0.05 for all comparisons).

Taxonomic Distribution Relative to the Sargasso Sea

For taxonomical distribution (Figure 8) at the domain level, Open Sound Puget Sound samples clustered together and in comparison relative to the Sargasso Sea, were more composed of Archaea and Eukaryota, with an underrepresentation of Bacteria. Nearshore Puget Sound samples, along with the WWTP effluent samples also clustered together (with the exceptions of the Hoodsport and the Marina 2010 sample), and were characterized by having higher levels of Bacteria and lower levels Archaea and Eukaryota than the Sargasso Sea sample. As we saw earlier with our LCA domain distributions, Nearshore samples have a higher relative abundance of Bacteria and lower relative abundances of Archaea and Eukaryota than open sound samples. Continuing to the phylum level, Puget Sound samples including the WWTP effluents follow the same pattern with an overrepresentation of Bacteroidetes and Actinobacteria (more specifically the samples that contained lower levels of salinity) and an underrepresentation of Proteobacteria and Cyanobacteria. Both WWTP effluent samples exhibited an over representation of Firmicutes, a phylum typically associated with wastewater that contains the fecal indicator bacteria fecal *Streptococci,* and *Enterococcus*.

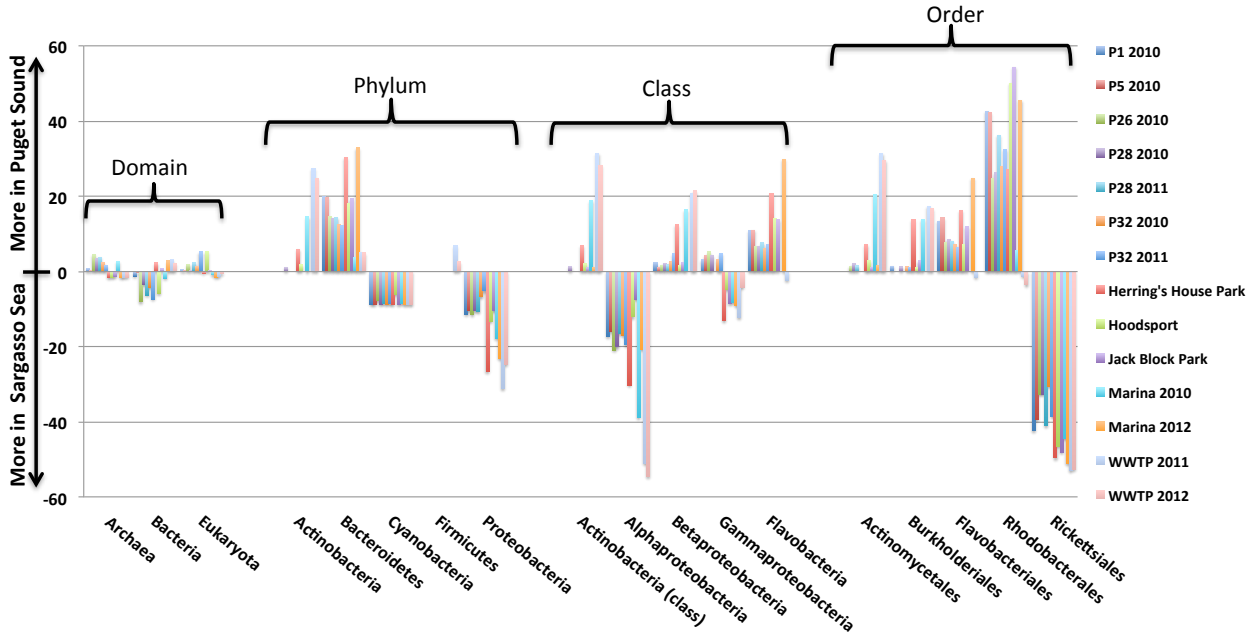

Figure 8. Predominant Taxonomic Distribution Relative to the Sargasso Sea. Only taxa and functions with a difference of proportions (percent difference in relative abundance between relative sample and Sargasso Sea) greater than 1% are shown. Positive values indicate higher relative abundance in Puget Sound and WWTP samples, and negative values indicate higher relative abundance in the Sargasso Sea.

Although the 2012 Puget Sound samples were dominated by Alphaproteobacteria at the class level, this graph depicts an underrepresentation of Alphaproteobacteria when compared to the Sargasso Sea, and an over representation of Betaproteobacteria. The class Gammaproteobacteria was split, with the Open Sound samples showing a higher abundance relative to the Sargasso Sea while the Nearshore and WWTP effluent samples exhibited a lower abundance than the Sargasso Sea. Gammaproteobacteria are generally considered to be more abundant in saline environments however, it appears they may have a threshold for salinity and may not be as tolerant as other saline acclimated classes. The relative abundance of Alphaproteobacteria in the majority of the Nearshore samples was far more underrepresented than in the 0pen Sounds samples when compared to the Sargasso Sea; verifying its preference for marine environments [47]. Betaproteobacteria

(more abundant in the Puget Sound and WWTP effluent) is commonly more numerous in freshwater, but given its overabundance in the WWTP effluent samples, this class may also be a taxonomic indication of a high plasticity taxa capable of acclimatizing and thriving in human impacted environments. This finding is further documented in the WWTP effluents, while their most abundant phylum, Proteobacteria was mainly composed of Betaproteobacteria, followed by Gammaproteobacteria and then Alphaproteobacteria. Overall, these outcomes suggest the diverse distribution of the most recognized phylum Proteobacteria, and the hierarchal composition may be associated with anthropogenic impact in addition to their preference for salinity.

At the order level, we continue to see a similar distribution of the class Alphaproteobacteria, with the Sargasso Sea being more abundant in Rickettsiales, while all the Puget Sound samples, especially those that are more saline, are more abundant in Rhodobacterales, a dominant order represented in both the 2010 and 2012 Puget Sound contigs [25]. The phylum Bacteroidetes, more abundant in the Puget Sound samples, seems to be partially driven at the class level by the overrepresented Flavobacteria, and at the order level by the overabundant Flavobacteriales. The orders, Actinomycetales, Burkholderiales were more abundant in the samples that had lower levels of salinity.

Functional Distribution Relative to the Sargasso Sea

Function distribution (Figure 9) in general was distributed by location, all Puget Sound samples and WWTP clustering together, with only a few exceptions. Thirteen of the twenty-eight SEED Level 1 functional categories (Cell Division and Cell Cycle, Dormancy and Sporulation, Fatty Acids, Lipids, and Isoprenoids, Iron acquisition and metabolism, Nitrogen Metabolism, Nucleosides and Nucleotides, Phosphorus Metabolism,

Photosynthesis, Potassium metabolism, Respiration, Secondary Metabolism, Stress Response, Sulfur Metabolism category) showed no difference greater than 1% between Puget Sound samples relative to the Sargasso Sea. Both WWTP samples followed the same trend except in the category "Nitrogen metabolism', where the WWTP 2012 effluent sample showed a slight overrepresentation and difference of proportion of approximately 1.4%. The remaining fifteen SEED Level 1 categories (including only samples with a difference of portions over 1%) showed a distinct separation in function based on Puget Sounds locations (including WWTP effluent samples) and the Sargasso Sea. The Sargasso Sea (Open Ocean) sample was more abundant in the "Amino Acids and Derivatives", "Carbohydrates", "Cell Wall and Capsule", "Clustering-based subsystems", "Cofactors, Vitamins, Prosthetic Groups", "Pigments", "DNA Metabolism", and "Miscellaneous" categories; with one exception in the category "Carbohydrates", where the WWTP effluent 2011 showed a higher representation by approximately 1.3%. The Puget Sound samples (Puget Sounds and WWTP effluent samples) showed an higher representation in SEED subsystem level 1 categories "Motility and Chemotaxis", "Membrane Transport", "Metabolism of Aromatic Compounds", "Phages, Prophages, Transposable elements, Plasmids". "Protein Metabolism", "Regulation and Cell signaling RNA Metabolism", and "Virulence, Disease and Defense".

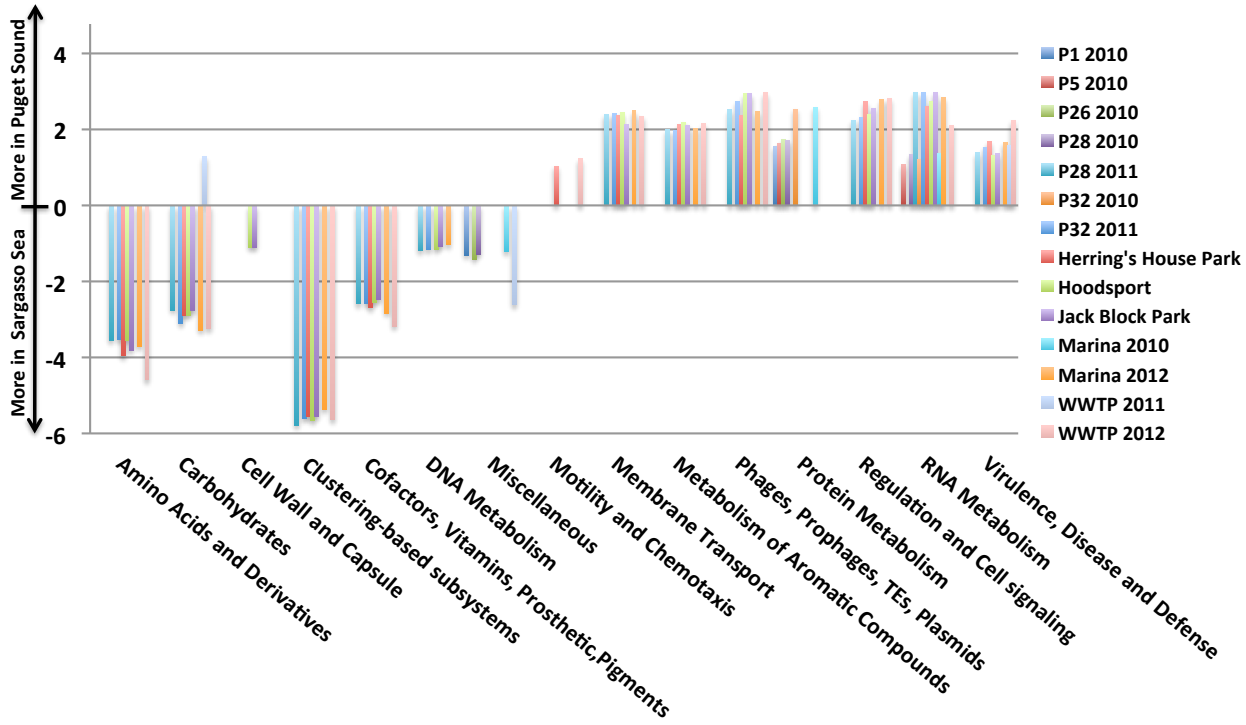

Figure 9. Predominant Functional Distribution Relative to the Sargasso Sea. Only taxa and functions with a difference of proportions (percent difference in relative abundance between relative sample and Sargasso Sea) greater than 1% are shown. Positive values indicate higher relative abundance in Puget Sound and WWTP samples, and negative values indicate higher relative abundance in the Sargasso Sea.

Earlier in our investigation, we decided to focus on categories with potential public health relevance with subsystems level 1 categories: “Phages, Prophages, Transposable elements, Plasmids”, “Stress Response” and “Virulence, Disease and Defense”. Two of three categories, “Phages, Prophages, Transposable elements, Plasmids” and “Virulence, Disease and Defense”, both affiliated with agents derived from human induced contamination (wastewater impact on environment) were more abundant in the Puget Sound and WWTP effluent samples when compared versus the Sargasso Sea. Upon closer review, both categories were primarily driven by the level 2 “Phages, Prophages “ and “Resistance to antibiotics and toxic compounds” (Supplementary Figure 7) as we saw earlier in the function distribution of the 2012 Puget Sound Samples. Overall, it appears that the

Sargasso Sea has presumably less human impact than the more coastal related areas (Open Sound and Nearshore) of the Puget Sound as exposed by the functional distribution of public health relevant categories further confirming an increased gradient of anthropogenic impact and health related consequences of coastal waters.

Species Diversity of Puget Sound Metagenome

In order to characterize bacterial species diversity in a community, we used the Shannon diversity index (H), an index that accounts for both the number of species in the community (richness) and their representative abundance (evenness)(Table 2). Indices were then converted to true diversity (effective number of species) for comparable interpretation of diversities from different communities by taking the equivalent number, the exponent of the Shannon diversity index (exp (H)) [41]. Equitability describes the evenness, calculated by Shannon's equitability (E_H), and assumes a value between zero and one, with one representing complete evenness.

Location	Species Richness (S)	Evenness (EH)	Diversity (exp H)
WWTP 12	1844	0.85	619
Marina 10	2014	0.82	530
P28 11	1702	0.82	462
p32 10	1754	0.81	435
P28 10	1758	0.81	425
P32 11	1694	0.81	423
P26	1739	0.81	422
P1	1734	0.79	373
WWTP 11	1796	0.78	356
P5	1998	0.77	348
HHP	2005	0.75	302
HOOD	1595	0.77	293
jbp	1520	0.76	262
Marina 12	1699	0.70	181

Table 2. Species Diversity of Puget Sound Metagenomes based on the exponential value of the Shannon Diversity Index and includes both Species Richness (S) and Species Evenness (EH).

Diversity indices ranged from 5.20-6.43 with the Marina 2012 sample being the least diverse with 180 equally-common species and the WWTP 2012 effluent sample being the most diverse with 618 equally-common species (approximately three times more diverse). Overall, the Nearshore samples were classified as the least diverse samples (181-302 equally-common species), and were closely trailed by the WWTP 2011 effluent then the Open Sound samples with the exception of the Marina 2010 (the second most diverse sample). The diversity of marine ecosystems is integral to their stability and function and our findings for our naturally sampled environments suggest the possible reduction of species diversity follows a gradient of anthropogenic contamination, a common factor associated with reductions in species richness and/or species evenness in marine habitats [9]. Equability values ranged from 0.70-0.85, again with the Marina 2012 sample with the least evenly distribute species abundance and the WWTP 2012 effluent with the most evenly distributed species abundances. In general, a community represented by species with equal abundances is more complex than a community with species with unequal abundances [54] as shown here, the WWTP sample did not have the most number of species but it did have the highest equability.

Although many marine habitats are inherently more diverse than others [9], using metagenomics to monitor coastal areas can inform us on the ecological composition and quality of a marine habitat. These studies will allow us to further explore the microbial community providing the tools to discern foundational species, microbial capabilities and environmental parameters associated with population dynamics. The coastal ecosystems are becoming increasingly burdened by the anthropogenic impact and contamination while subsequently human-health related concerns and safety are becoming more significant.

Polluted areas, such as the Lower Duwamish River (Herring's House Park sample), a Superfund site, are areas of increased interest and concern since these environments are considered to be more vulnerable [6, 9] to the harmful effects of global climate changes. As temperatures continue to rise, temperature increases will interact with pollution, and some contaminants are predicted to increase in toxicity and bioavailability [55]. As a result, contaminated environments with temperature induced pollutants should be given a high priority for remediation, as these areas can be considered less resilient and more prone to ecosystem destruction [56].

Environmental Metadata

Relationships between taxonomic composition and metadata were assessed at the Phylum level using linear regression. Only metadata collected across all samples were examined and included temperature and salinity (Supplementary Table 1, 3). All taxa were normalized (relative abundance) prior to performing pairwise correlations (Pearson). As identified in the Puget Sound foundations study [25] a negative correlation between Actinobacteria abundance and salinity ($r= -0.701$; $p<0.011$) was found across all Puget Sound metagenomes ($r= -0.939$; $p<0.000$ when WWTP effluent samples were included) (Supplementary Figure 8). These finding further substantiate earlier results and previous findings of the freshwater and terrestrial signals associated with Actinobacteria abundance [57, 58]. Additionally, a negative correlation was found between Firmicutes abundance and salinity across Puget Sound samples when WWTP effluent samples were included ($r= -0.793$; $p<0.001$) and a positive correlation between Proteobacteria abundance and salinity across Puget Sound samples ($r= 0.816$; $p<0.001$) and when WWTP effluent samples were included ($r= 0.869$; $p<0.000$) (Supplementary Figures 9, 10 respectfully). Bacteroidetes

abundances were positively correlated with temperature across all Puget Sound samples ($r= 0.618$; $p<0.032$) (Supplementary Figure 11) and results were clearly seen by the increased relative abundance of Bacteroidetes in the Marina 2012 sample from the 2010 sample (temperature increase ~ 7° C). Firmicutes abundances were negatively correlated with temperature across all Puget Sound samples ($r= -0.713$; $p<0.009$) (Supplementary Figure 12). Lastly, among predominant taxa, there was a negative correlation found between Proteobacteria and Bacteroidetes abundances ($r= -0.624$;$p<0.030$) (Supplementary Figure 13).

Repeat Sample Analyses

One of the challenges for metagenomic monitoring is the ability and feasibility to produce high quality data with replicable results in real time. As seen in Figure 10, our results verify the high reproducibility and discriminatory capabilities of metagenomic profiling when using a quality driven, comparable methods.

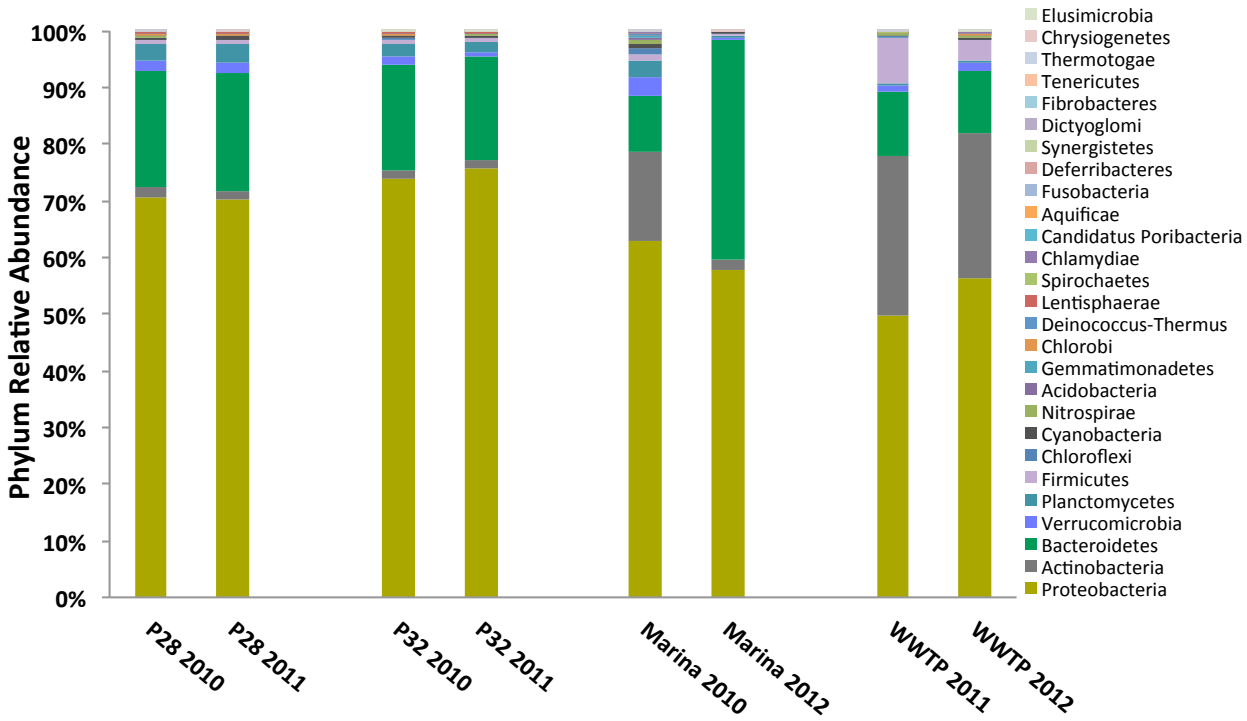

Figure 10. Phylum Distribution of Repeat Samples with LCA annotation using the MG-RAST Pipeline. Phylum abundance counts were normalized to the total number of Phylum abundance counts per sample.

Repeat samples in the Open Sound, taken approximately a year apart exhibited highly similar composition, while repeat samples taken from the Marina and WWTP effluent during different seasons displayed considerable compositional differences. These results suggest that environmental conditions (in addition to other unidentified variables) influence taxonomic species richness and evenness (diversity) (Supplementary Table 1, 3). Environmental variables that varied significantly between the Marina and WWTP effluence include both temperature and salinity as seen in Figure 11.

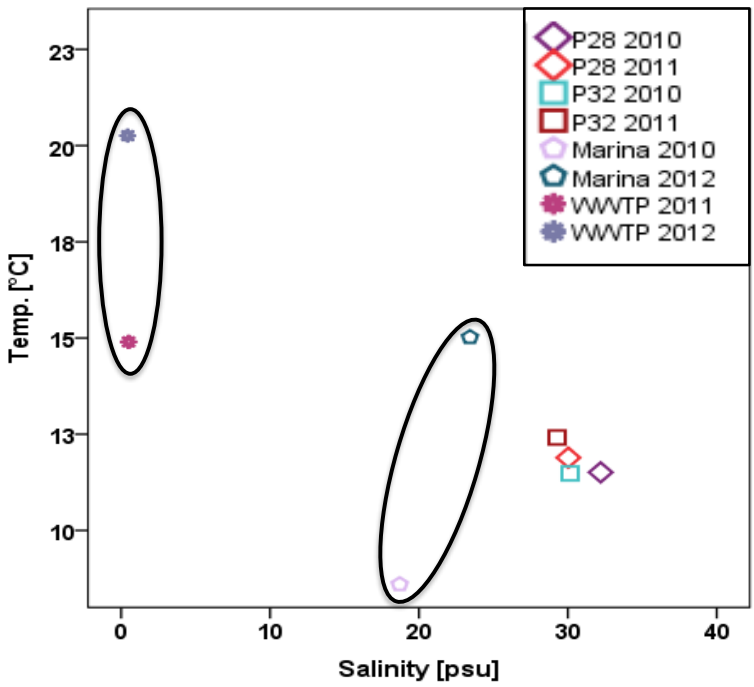

Figure 11. Environmental Determinants of Compositional Diversity of Repeat Samples. Open Sound samples station P28 and P32 both have similar environmental conditions (temperature and salinity) as seen by the clustering and both exhibit similar compositional richness and evenness at the phylum level annotation. The Marina and WWTP effluent however had different levels of temperature and salinity, and are showing different cluster presentations as well as phylum compositions.

Figure 11 shows how the environmental conditions of temperature and salinity, two of the most important environmental parameters, influence the bacterial diversity seen at the phylum level composition as exhibited in Figure 10 [59]. Repeat Open Sound samples, which both were taken under similar environmental conditions are exhibiting highly similar phylum level composition while repeat samples Marinas and WWTP effluents taken under different environmental conditions are not clustering together and are exhibiting significantly different phyla presentations. Typically, as a result of environmental patchiness it is difficult to take "true replicates"[60], but our longitudinal study shows the reproducibility and discernibility of using 454/Roche Pyrosequencing. These results further validate the scope and successful applications of metagenomic studies.

Antibiotic Resistance Determinants Index

In this study we used the broad scale screening approach of the Antibiotic Resistance (ARD) Index established by Port et al., in "Metagenomics and Antibiotic Resistance surveillance" to look for public health relevant antibiotic resistant determinants. The ARD index is a highly sensitive decision tool and is based on the method of using metagenomics in combination with 454-pyrosequencing and bioinformatics analysis to survey antibiotic resistance determinants from intact marine communities and WWTP genomes. Currently, environmental health monitoring is centered on the culture-based methods of identifying single indicator organisms that have been historically associated with harmful human pathogens. Putative levels of ARD in the environment were assessed using the following determinant categories: gene transfer potential, antibiotic resistance genes and pathogenicity potential using the following criteria shown in Table 3.

Index	Criteria
Antibiotic Resistance Gene Potential	
Antibiotic Resistance Genes	80% ID; 50 aa
Metal Resistance Genes	60% ID; 50 aa
Gene Transfer Potential	
Plasmids	95 % ID; 100 bp
Transposable Elements	80% ID; 50 aa
Phages	60% ID; 50 aa
Pathogenicity Potential	
LCA Genus Pathogens	95% ID; 50 aa

Table 3. Antibiotic Resistance Determinant Index with Criteria.

Antibiotic Resistance Gene Potential

Overall, 48 Antibiotic resistance genes (ARGs) (0.0020% of all qc sequence reads) were identified using the ARDB+ with a minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids. The WWTP effluent samples

had the highest representation (33), followed by the Nearshore (13) and Open Sound (2). Bacitracin resistance genes were most abundant accounting for 22% of ARGs, followed by Aminoglycoside (19%), Macrolide (15%), Tetracycline (10%) and Beta-lactam (8%). Metal resistance genes (MRGs) were identified through the functional annotations of the SEED subsystems hierarchal classification system via MG-RAST pipeline [38]. All genes coding for metal resistance in a subset of the Level 2 "Resistance to antibiotics and toxic compounds" with an e-value threshold of 1.0 e-5, a minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids were retained. Overall, the category "Cobalt-zinc-cadmium resistance" accounted for approximately 60.4% of MRGs annotations, followed by "Arsenic resistance" (22%), "Mercury resistance operon" (14%), Zinc resistance (3.5%), and Cadmium resistance (0.1%). MRGs relative abundances ranged form 0.74% to 0.09% of total annotated functional genes, with the highest representation in the Marina 2012 sample and lowest in P5 2010, and P6 2010. Location averaged relative abundances were highest in the WWTP effluent samples (0.61%), then the Nearshore samples (0.56%), and were lowest in the Open Sound samples (0.25%).

Gene-Transfer Potential

A total of 1205 Plasmid sequences (0.60% of total sequence reads) were identified using the NCBI RefSeq plasmid database applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 95% identity, and a minimum alignment length of 50 base pairs. The WWTP effluent had the highest representation, accounting for 70% of the reads, followed by the Nearshore samples (20%) and the Open Sound (10%). Overall, plasmid sequences were affiliated with 83 genera and 189 different species and predominately represented by the genera Acinetobacter 15%, Ruegeria 12% and

Paracoccus, Lactococcus, and Pseudomonas all at 5%. Approximately 37% of the plasmid sequences were considered to be pathogenic and was represented by the following genera (Phylum): Acinetobacter 41%, Pseudomonas 13%, Klebsiella 10%, Escherichia 7%, (all from Gammaproteobacteria), Ralstonia 6% (Betaproteobacteria), and Enterococcus 6% (Firmicutes). Pathogenic plasmid sequences were distributed mainly in the WWTP effluents 89%, Nearshore 9% (Marina, Herring's House Park and JBP) and the Open Sound 2% (stations P28, P32). In total, 4089 Transposable elements (TEs) (~0.17% of all qc sequence reads) were identified through GenBank applying a minimum sequence similarity criteria of 80%, and a minimum alignment length of 50 amino acids. The WWTP effluent had the highest representation, accounting for 69% of the reads, followed by the Nearshore samples (20%) and the Open Sound (11%.) The most represented genera across all reads included Octadecabacter, Thalassiobium and Rhodobacteraceae all at around 10%. Pathogenic genera accounted for ~18% of total reads and were distinguished by 26 genera with the highest representation Acinetobacter (21%), Mycobacterium (13%), Pseudomonas (12%) and Klebsiella, Streptococcus, and Escherichia around 5% each. The phylum Gammaproteobacteria (Acinetobacter, Pseudomonas, Klebsiella, and Escherichia) was most abundant across pathogenic TEs. Pathogenic plasmid sequences were distributed mainly in the WWTP effluent (72%), followed by the Nearshore (19%), and the Open Sound (9%). Phages were taxonomically annotated through the MG-RAST pipeline and included all unassembled sequence reads matching to the Virus domain using an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 80% identity, and a minimum alignment length of 50 amino acids. Overall, Nearshore samples had the highest

averaged relative abundance with Phages accounting for ~1.2% of annotated domains, followed by Open Sound (0.97%) and WWTP effluents (0.59%).

Pathogenicity Potential

Pathogens were taxonomically identified by comparing the genus level annotations from the Microbial Rosetta Stone Database [24] to the genus level based annotations using the LCA applying an e-value threshold of 1.0e-5, a minimum sequence similarity criteria of 95% identity, and a minimum alignment length of 50 amino acids. The WWTP effluent accounted for 54% of total pathogenic reads, the Nearshore 35% and the Open Sound 11%. The most potential pathogenic genera included: Acinetobacter (14%), Clostridium (13%), Streptococcus 11%), Vibrio and Bacillus (each at 10%).

Antibiotic Resistance Determinants Index Summary

Using population level surveillance, we examined community relative abundances of putative opportunistic pathogens and resistance genes across all Puget Sound and WWTP metagenomes (Supplementary Table 4). Samples were grouped according to location (Open Sound, Nearshore, WWTP) and preliminary results exposed an increase in averaged relative abundance across a gradient of anthropogenic impact between the Open Sound, Nearshore and WWTP effluent samples in all AR determinants, with the exception of the sub-category Phages which exhibited highest representation in Nearshore samples (Figure 12). The trendline in Figure 12 illustrates the results of the averaged relative abundance of the determinants samples grouped by location from Open Sound to Nearshore to the WWTP.

	Open Sound	Nearshore	WWTP	Trendline
ARGs	0.0002%	0.0010%	0.0145%	
MRGs	0.25%	0.56%	0.61%	
Plasmids	0.01%	0.02%	0.36%	
TEs	0.09%	0.16%	0.57%	
Phages	0.97%	1.20%	0.59%	
Pathogens	0.0011%	0.0029%	0.0217%	

ARG= Antibiotic Resistant Genes; MRG= Metal Resistant Gens; TE: Transposable Elements

Figure 12. Antibiotic Determinant Index Results by Location (Open Sound, Nearshore, WWTP). Using the average relative abundance: all data increases from Open Sound samples to Nearshore to WWTP, except for the subcategory Phages where Nearshore samples have the highest mean relative abundance.

CONCLUSION

As environmental sequencing technologies become more affordable and resourceful, additional opportunities for metagenomics applications develop. These studies over time will disclose vast amounts of information on novel organisms and metabolic processes allowing scientists to better characterize microbial phenotypes and environmental niches, signifying key species and roles fundamental in sustaining an efficient, functional environment. In addition to its ecological relevance, advanced metagenomic surveillance of environments allows for translational research into human health by increasing the limits of detection and quantification for the evolution and progression of human health determinates.

Study limitations of using metagenomic applications were discussed throughout the paper and briefly include cost and the complexity of bioinformatics analysis and annotation

limits due to taxonomic classification inconsistencies, scarcity of reference genomes (bias towards cultured organisms), and limits on sequencing length and depth resulting in dominant populations overshadowing low abundance taxa and lower annotation specificity. Although challenges exist, the discovery and benefits of using metagenomic studies greatly compensates for the limitations. Future metagenomic applications will only evolve and studies will continue to be successful in the characterization and remediation of our most influential environments.

The objective of this study was to further define the Puget Sound metagenome by continuing to monitor and build upon the taxonomic composition, functional potential and comparative analysis of the surface water bacterial communities along the coastal areas of the Puget Sound and to more effectively address the potential environmental signals of human and biological impact that may be relevant to human health. Our results revealed the high reproducibility and discriminatory capabilities of metagenomic profiling and determined high throughput metagenomic analysis is an appropriate method to assess the taxonomic characterization and functional potentials of microbial populations. Advanced characterization of environments through annotated metagenomic data and marine associated metadata can help define the physical, chemical and biological parameters that play a contributing role in the community diversity and function necessarily to sustain a flourishing environment. Comparative analysis of all metagenomes exposed significant differences in both microbial diversity and human health determinants across a gradient of anthropogenic impact further illustrating human impacts on marine ecosystems and marine impacts on human health. Overall, our results demonstrated our improved

characterization of the Puget Sound, as well as the future applications and significance of metagenomic analyses in environmental health monitoring and surveillance.

Metagenomics in combination with next generation sequencing and bioinformatics offers an unprecedented, sensitive approach to evaluate intact community genomes for the novel detection and characterization of microbial populations. Its gene-based, population level surveillance, provides advanced insight into uncultured organisms broadening understanding of environments, their indigenous microbial compositions and their functional potential. This data will further initiate longitudinal monitoring of Puget Sound providing innovative, translational research to further characterize and expand environmental monitoring and public health impact and awareness.

REFERENCES

1. Karl, D.M., *Nutrient dynamics in the deep blue sea.* Trends Microbiol, 2002. **10**(9): p. 410-8.
2. Gianoulis, T.A., et al., *Quantifying environmental adaptation of metabolic pathways in metagenomics*, in *Proc Natl Acad Sci U S A*. 2009: United States. p. 1374-9.
3. Kisand, V., et al., *Phylogenetic and functional metagenomic profiling for assessing microbial biodiversity in environmental monitoring*, in *PLoS One*. 2012: United States. p. e43630.
4. Amann, R.I., W. Ludwig, and K.H. Schleifer, *Phylogenetic identification and in situ detection of individual microbial cells.* Microbiol Rev, 1995. **59**(1): p. 143-69.
5. Herr, D. and R.G. Galland, *The Ocean adn Climate Change. Tools and Guidelines for Action*. 2009 IUCN: Gland, Switzerland. p. 72
6. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: the unseen majority.* Proc Natl Acad Sci U S A, 1998. **95**(12): p. 6578-83.
7. Wilson, D.J., *Insights from genomics into bacterial pathogen populations*, in *PLoS Pathog*. 2012: United States. p. e1002874.
8. Levinton, J.S., et al., *Rapid loss of genetically based resistance to metals after the cleanup of a Superfund site*, in *Proc Natl Acad Sci U S A*. 2003: United States. p. 9889-91.
9. Johnston, E.L. and D.A. Roberts, *Contaminants reduce the richness and evenness of marine communities: a review and meta-analysis*, in *Environ Pollut*. 2009: England. p. 1745-52.
10. Crossett, K., et al., *National Coastal Populations Report. Population Trends from 1970-2020.*, in *NOAA's State of the Coast*. 2013 National Oceanic and Atmospheric Administration.
11. de Forges, B.R., J.A. Koslow, and G.C. Poore, *Diversity and endemism of the benthic seamount fauna in the southwest Pacific.* Nature, 2000. **405**(6789): p. 944-7.
12. Howarth, R., et al., *Nutrient Pollution of Coastal Rivers, Bay and Seas.* Issues in Ecology, 2000 (7): p. 15
13. Varela, A.R. and C.M. Manaia, *Human health implications of clinically relevant bacteria in wastewater habitats.* Environ Sci Pollut Res Int, 2013.
14. Henze, M., et al., *Biological wastewater treatment principles, modelling and design*. 2008 London: IWA.
15. Tchobanoglous, G., F. Burton, and H. Stensel, *Wastewater engineering (treatment disposal reuse)* 4th McGraw-Hill ed. 2003 New York: Metcalf & Eddy Inc.
16. Baquero, F., J.L. Martinez, and R. Canton, *Antibiotics and antibiotic resistance in water environments.* Curr Opin Biotechnol, 2008. **19**(3): p. 260-5.
17. Gootz, T.D., *The global problem of antibiotic resistance.* Crit Rev Immunol, 2010. **30**(1): p. 79-93.
18. Martinez, J.L., *Antibiotics and antibiotic resistance genes in natural environments*, in *Science*. 2008: United States. p. 365-7.
19. Zhang, T., X.X. Zhang, and L. Ye, *Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge*, in *PLoS One*. 2011: United States. p. e26041.

20. Frost, L.S., et al., *Mobile genetic elements: the agents of open source evolution*, in *Nat Rev Microbiol*. 2005: England. p. 722-32.
21. Munoz-Lopez, M. and J.L. Garcia-Perez, *DNA transposons: nature and applications in genomics.* Curr Genomics, 2010. **11**(2): p. 115-28.
22. Kristensen, D.M., et al., *New dimensions of the virus world discovered through metagenomics.* Trends Microbiol, 2010. **18**(1): p. 11-9.
23. Lohr, J.E., F. Chen, and R.T. Hill, *Genomic analysis of bacteriophage PhiJL001: insights into its interaction with a.* Appl Environ Microbiol, 2005. **71**(3): p. 1598-609.
24. Ecker, D.J., et al., *The Microbial Rosetta Stone Database: a compilation of global and emerging.* BMC Microbiol, 2005. **5**: p. 19.
25. Port, J.A., et al., *Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound*, in *PLoS One*. 2012: United States. p. e48000.
26. Babson, A., A. Kawase, and P. MacCready, *Seasonal and interannual variability in the circulation of Puget Sound, Washington: A box model study* 2006 Atmos Ocean.
27. Seattle, C.o. *Protecting Seattle's Waterways*. 2013 [cited 2013 April 15]; http://www.seattle.gov/util/EnvironmentConservation/Projects/DrainageSystem/SewageOverflowPrevention/index.htm%5D.
28. ***Restoration Activities Case: Elliott Bay/Duwamish River, WA*** 2013 [cited 2013 April, 10]; http://www.darrp.noaa.gov/northwest/elliott/seabd.html%5D. Available from: http://www.darrp.noaa.gov/northwest/elliott/seabd.html.
29. EPA, U.S., *Five-Year Review Report Pacific Sound Resources Superfund Site Seattle, King County, Washington*. 2004 Region 10: Washington p. 224.
30. *Beach Access Opening Celebration at Jack Block Park*, in *West Seattle Herald*. 2011: Washington.
31. Georgeson, A., W. Matthews, and P. Orth, *Hood Canal Pollution Identification and Correction Project Final Report*. 2008: Mason County Public Health. p. 58.
32. County, K. *West Point Treatment Plant*. 2012[cited 2013 March 15]; http://www.kingcounty.gov/environment/wtd/About/System/West/Process.aspx%5D.
33. Gilbert, J.A. and C.L. Dupont, *Microbial metagenomics: beyond the genome.* Ann Rev Mar Sci, 2011. **3**: p. 347-71.
34. Budnick, G., R. Howard, and D. Mayo, *Evaluation of Enterolert for Enumeration of Enterococci in Recreational Water. Applied and Environmental Microbiology* 1996 (62): p. 3881–3884.
35. Ecology, W.S.D.o. *Beach Environmental Assessment and Communication and Health*. [cited 2013 April 23]; http://www.ecy.wa.gov/programs/eap/beach/index.html%5D.
36. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic.* BMC Bioinformatics, 2008. **9**: p. 386.
37. *Genomic Standards Consortium*. 2013 [cited 2013 April 15]; http://gensc.org/gc_wiki/index.php/Main_Page%5D.
38. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes*, in *Nucleic Acids Res*. 2005: England. p. 5691-702.

39. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*, in *Appl Environ Microbiol*. 2007: United States. p. 5261-7.
40. Wooley, J.C., A. Godzik, and I. Friedberg, *A primer on metagenomics.* PLoS Comput Biol, 2010. **6**(2): p. e1000667.
41. Jost, L., *Entropy and diversity.*
42. Zhu, W., A. Lomsadze, and M. Borodovsky, *Ab initio gene identification in metagenomic sequences.* Nucleic Acids Res, 2010. **38**(12): p. e132.
43. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
44. Parks, D.H. and R.G. Beiko, *Identifying biologically relevant differences between metagenomic communities*, in *Bioinformatics*. 2010: England. p. 715-21.
45. Gomez-Alvarez, V., T.K. Teal, and T.M. Schmidt, *Systematic artifacts in metagenomes from complex microbial communities*, in *ISME J*. 2009: England. p. 1314-7.
46. Niu, B., et al., *Artificial and natural duplicates in pyrosequencing reads of metagenomic data.* BMC Bioinformatics, 2010. **11**: p. 187.
47. Newton, R.J., et al., *A guide to the natural history of freshwater lake bacteria.* Microbiol Mol Biol Rev, 2011. **75**(1): p. 14-49.
48. Pereira, F., et al., *Identification of species by multiplex analysis of variable-length sequences*, in *Nucleic Acids Res*. 2010: England. p. e203.
49. Kolbert, C.P. and D.H. Persing, *Ribosomal DNA sequencing as a tool for identification of bacterial pathogens*, in *Curr Opin Microbiol*. 1999: England. p. 299-305.
50. Nematollahi, A., et al., *Flavobacterium psychrophilum infections in salmonid fish.* J Fish Dis, 2003. **26**(10): p. 563-74.
51. *Recreational Water Quality Criteria*, U.S. EPA, Editor. 2012. p. 69.
52. Washington, D.o.E.S.o. *Surface Water Criteria*. **Table 210 (3)(b) Water Contact Recreation Bacteria Criteria in Marine Water** 2007 [cited 2013 March 10]; http://www.ecy.wa.gov/programs/wq/swqs/criteria-marine/wac173201a_210-bacteria.html%5D.
53. Rusch, D.B., et al., *The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through.* PLoS Biol, 2007. **5**(3): p. e77.
54. Kunin, V., et al., *A bioinformatician's guide to metagenomics.* Microbiol Mol Biol Rev, 2008. **72**(4): p. 557-78, Table of Contents.
55. Sokolova, I.M. and G. Lannig.
56. Hooper, D.U., et al.
57. Aguilo-Ferretjans, M.M., et al., *Phylogenetic analysis of the composition of bacterial communities in human-exploited coastal environments from Mallorca Island (Spain).* Syst Appl Microbiol, 2008. **31**(3): p. 231-40.
58. Kelly, K.M. and A.Y. Chistoserdov, *Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island).* FEMS Microbiol Ecol, 2001. **35**(1): p. 85-95.
59. Lozupone, C.A. and R. Knight, *Global patterns in bacterial diversity.* Proc Natl Acad Sci U S A, 2007. **104**(27): p. 11436-40.
60. Luo, C., et al., *Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample.* PLoS One, 2012. **7**(2): p. e30087.