Spring 2023

A Comparative Metagenomics Study on A Bioreactor System in Salinas, CA, The Salinas River Valley, and The Tijuana River Valley

Connie Samantha Machuca

A Comparative Metagenomics Study on a Bioreactor System in Salinas, CA, The Salinas River Valley, and The Tijuana River Valley

A Thesis

Presented to the

Faculty of the

Department of Applied Environmental Science

California State University Monterey Bay

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Environmental Science

by

Connie Samantha Machuca

Term Completed: Spring 2023

CALIFORNIA STATE UNIVERSITY MONTEREY BAY

The Undersigned Faculty Committee Approves the

Thesis of Connie Samantha Machuca:

A COMPARATIVE METAGENOMICS STUDY ON A BIOREACTOR SYSTEM IN SALINAS, CA, THE SALINAS RIVER VALLEY, AND THE TIJUANA RIVER VALLEY

Nathan Jue

Nathaniel K. Jue, PhD
Department of Biology and Chemistry
California State University, Monterey Bay

Arlene Haffa (May 17, 2023 07:36 PDT)

Arlene Haffa, PhD
Department of Biology and Chemistry
California State University, Monterey Bay

Kelly Goodwin - NOAA Federal
Kelly Goodwin - NOAA Federal (May 17, 2023 11:19 PDT)

Kelly Goodwin, PhD
National Oceanic and Atmospheric Administration
Miami, FL

Doug Smith

Doug Smith, Dean (Interim)
Dean of Graduate Studies and Research

22 May 2023
Approval Date

ABSTRACT

A Comparative Metagenomics Study on A Bioreactor System
in Salinas, CA, The Salinas River Valley, and The Tijuana
River Valley
by
Connie Samantha Machuca
Master of Science in Environmental Science
California State University Monterey Bay, 2023

Coastal environments are some of the most productive and valuable ecosystems in the world while also having some of the highest levels of non-point source pollution. Genetic analyses of bacteria have provided scientists with a better understanding of how pollution affects functional potential within the environment. This is done by evaluating the presence/absence of particular taxonomic classifications as well as genes and gene functional groups. This study aimed to use genetic information isolated from bacteria found in two coastal watersheds, as well as a field bioremediation system, to learn how bacterial diversity and taxonomic groups differ between locations, the potential of sampled environments to remediate pollutants, and which functional groups are significantly represented within different locations. Amplicon sequencing of the 16S rRNA gene as well as whole metagenome shotgun sequencing were performed using isolated DNA from sediment samples collected from sites in two coastal watersheds, the Salinas River Watershed and the Tijuana River Watershed. Amplicon sequencing showed significant differences in alpha and beta diversity within different location sites. Beta diversity was also observed to be significantly affected by various environmental variables within location sites. Whole metagenome shotgun sequencing produced 60 high-quality dereplicated metagenomically assembled genomes (MAGs). MAGs from two location sites were found to have all genes necessary to complete two functional KEGG pathways related to agricultural runoff reduction. Hierarchical clustering of sequences within the high-quality dereplicated MAGs was also observed revealing over and under representation of 7 and 19 Level-3 GO categories, respectively. The genetic properties of bacteria found within this study's sampling locations provides local policymakers with information related to an ongoing bioremediation project as well as the function of the ecosystems that are vital for the regional and national economy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my committee members Dr. Jue, Dr. Haffa, and Dr. Goodwin for their invaluable guidance throughout my time completing my thesis project. I would also like to recognize Dr. Nastassia Patin for her extensive support and technical expertise throughout the data analysis portion of my project. Thank you to the NOAA AOML lab, specifically Dr. Goodwin, Dr. Patin, Dr. Thompson and Dr. Lim, for hosting me as a summer intern in 2022. My internship with NOAA allowed me to make significant progress on the bioinformatics portion of my project. Thank you to the WEE Lab at CSUMB and the Elkhorn Slough National Estuarine Research Reserve for providing me with the technology and supplies needed to complete the field work portion of my project. I am thankful for the Jue Lab, specifically Rahil Ryder, Bryan van Orman, Timothy Randazzo, Vanessa Garcia, Malia Machado, Nate Walton, Avalon Surratt, Sam Manges, Adam Dean, Stephanie Faris, and Mary Snook for their tireless effort during field days and sample processing in the laboratory.

This work was funded by the NOAA Center for Coastal and Marine Ecosystems Graduate Fellowship (NA16SEC4810009), the CSU Program for Education and Research in Biotechnology Faculty-Graduate Student Research Collaboration Program (CSUPERB 2021-2022), the CSU Agricultural Research Institute, and the California Leafy Greens Research Program.

Introduction

Coastal zones are known to be some of the most productive and valuable ecosystems in the world (Chouinard et al. 2015). In many coastal watersheds located in California, non-point sources of pollution have led to negative effects on human health and the environment (Stein et al. 2016; Dowd et al. 2008; Shrestha et al. 2020). Some negative consequences of coastal pollution include loss of habitat and biodiversity of living organisms, a shift in environmental processes, and a decline in ecosystem services (Nunes & Leston 2020). One way to assess environmental pollution levels and their magnitude is by studying microbes found within these areas. Microbes found within coastal environments are responsible for various biogeochemical processes that are necessary for the environment to function properly (Vincent et al. 2021). In addition, microbes are sensitive to changes within the environments they are found in, making them an appropriate study subject for environmental monitoring (Zhang et al. 2016).

Studying bacteria found within the environment has proven difficult in the past with only about 2% of all bacteria being culturable within a laboratory (Wade 2022). Metagenomics is a term used to describe the study of uncultured microorganisms at the genomic level (Wooley et al. 2010). High-throughput sequencing technology complemented with metagenomics approaches have made studying non-culturable bacterial species found within the environment more attainable (Wooley et al. 2010; Pérez-Cobas et al. 2020). Two primary methods used to study microbial communities in different capacities with the use of high-throughput sequencing are marker gene amplification and whole metagenome shotgun sequencing (shotgun metaG) (Pérez-Cobas et al. 2020). Marker gene amplification involves sequencing a region of a genome associated with a specific gene (Pérez-Cobas et al. 2020). Within bacterial studies, this specific gene of interest is the 16S rRNA gene that consists of 9 hypervariable regions and has a sequence length of about 1,550 bp long (Clarridge 2004; Janda et al. 2017). Amplification of the 16S rRNA gene has been used to study the diversity and complete taxonomic resolution of microbial communities in the environment (Fadeev et al. 2021). Shotgun metaG is a more comprehensive analysis since it provides sequences that are able to be assembled into entire genomes and can inform functional inferences associated with an environment based on gene presence/absence (Sharpton 2014). Both these approaches have led to various advancements in microbial ecology and diversity (Janda et al. 2017; Pérez-Cobas et al. 2020; Sharpton 2014).

In addition to using microbes as a means of environmental monitoring, they have also been used to detoxify and remediate pollutants (Chandran et al. 2020). Bioremediation involves using microorganisms to decrease or detoxify environmental contaminants to non-toxic levels (Raffa & Chiampo 2021; Sharma 2012; Mueller et al. 1996). Bioremediation is often done in a controlled environment with the use of bacteria that transform unwanted contaminants during several metabolic processes. Researchers at California State University, Monterey Bay (CSUMB) have been using a bioremediation approach to reduce contaminants from local waterways.

The Jue and Haffa Labs at CSUMB received funding from the California Department of Pesticide Regulation, the California State University Agricultural Research Institute, and the California Leafy Greens Research program to research the implementation of bioremediation systems in agricultural areas near Salinas, California. This research has identified potential bacterial pesticide remediators that would later be added to an experimental bioremediation system to remove unwanted contaminants from agricultural runoff. This bioreactor consists of a water-flow structure with a woodchip-filled trench to remove unwanted contaminants from agricultural runoff. Preliminary data from the bioreactor system in Salinas, California showed an overall reduction in pesticide abundance in imidacloprid and pyrethroid pesticides and significant changes in bacterial community composition and species diversity over time.

Two remediation pathways of interest that have been related to the reduction of agricultural pollutants are denitrification and dissimilatory nitrate reduction to ammonia (DNRA). Denitrification is the process of removing nitrate from the environment through plant, algae, and microbe absorption that involves oxidation of ammonia (Tomasek et al. 2017; Shapleigh 2006; Kuenen, 2008). Denitrification specifically involves reducing nitrate to gaseous nitrogen with the creation of nitrite, nitric oxide, and nitrous oxide where oxygen is limited (Tomasek et al. 2017). Some microbes perform DNRA in place of denitrification to complete respiration by using nitrate as an electron acceptor in place of oxygen (Kamp et al. 2015; Thamdrup 2012). In addition to reducing nitrate, DNRA is also responsible for converting nitrate to ammonia and allows for nitrogen to stay within the environment for other organisms to use (Bu et al. 2017). The steps involved in DNRA include nitrate reduction to nitrite and nitrite reduction to ammonia (Bu et al. 2017).

In this study, two different genetic approaches were used to provide information on the state of historically polluted environments and their metabolic potential to remediate pollutants. First, DNA was extracted from soil samples collected from two polluted coastal watersheds, a bioreactor system, and two comparison locations within the same areas. This was followed by 16s rRNA sequencing and whole metagenome shotgun sequencing (shotgun metaG), leading to high-quality amplicon sequence variants (ASVs) and metagenomically assembled genomes (MAGs) which were analyzed using comparative genomics, including alpha and beta diversity, taxonomic classifications, the presence of genes associated with two metabolic pathways, and Gene Ontology annotations.

Methods

Study Area(s)

The sampling regime for this study included three distinct locations: the Mexican/U.S.A border, the Salinas River Valley, and a constructed woodchip field bioreactor established near an agricultural field about a 1-mile distance from the Salinas River.

Mexican/U.S.A Border

Three sites near the Tijuana River Valley were sampled for this study: Smuggler's Gulch, Goat Canyon, and Los Peñasquitos Canyon Preserve – Canyonside Creek (Figure 1). Smuggler's Gulch and Goat Canyon have been used as monitoring sites in past studies to collect environmental metadata as well as to observe changes in raw sewage and heavy metal contamination (U.S. Customs and Border Protection n.d.). Smuggler's Gulch and Goat Canyon are categorized as "polluted Tijuana" sampling sites within this research study. Los Peñasquitos Canyon Preserve is owned and operated by both the City and County of San Diego. Los Peñasquitos Canyon Preserve is found within the Los Peñasquitos Watershed Management Area (WMA), which is primarily used for residential purposes, transportation, industry, and agriculture. Because of its location, Los Peñasquitos Canyon Preserve is assumed to be impacted by pollutants from urban sources. Within this study, Los Peñasquitos Canyon Preserve is categorized as a "comparison Tijuana" sampling site.

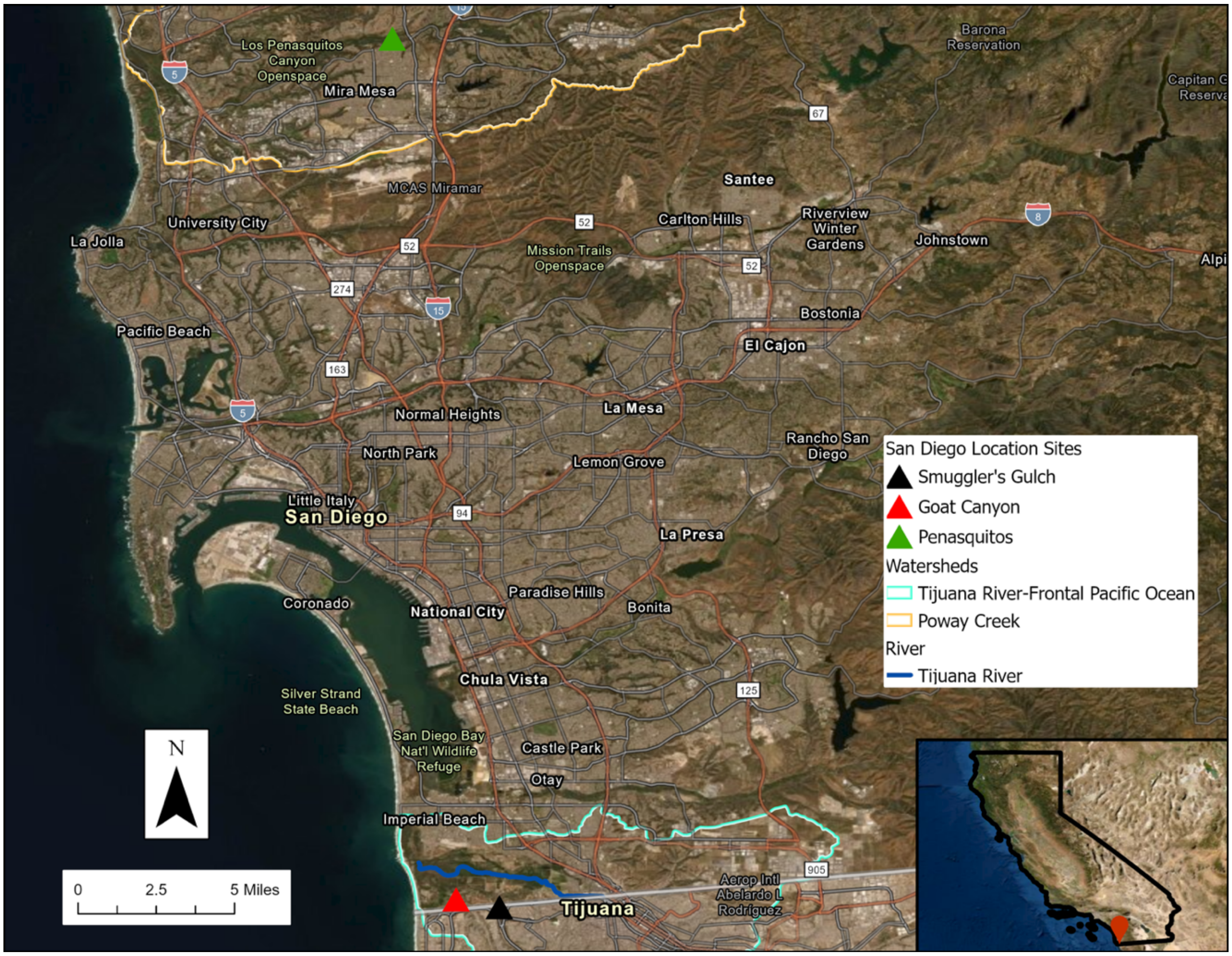

Figure 1. Map of the sampling sites near the Tijuana River Valley.

Salinas River Valley

Three sites near the Salinas River Valley were sampled for this study: CDPR Monitoring Site (Salinas River at Davis Rd), Salinas River National Wildlife Refuge (SRWR), and Arroyo Seco (Figure 2). The SRWR site is managed by the U.S. Fish and Wildlife Service and is located where the Salinas River empties into the Monterey Bay. Both the CDPR monitoring site and the SRWR are categorized as “polluted Salinas” sampling sites. Arroyo Seco is located about 40 miles from the Salinas River and flows

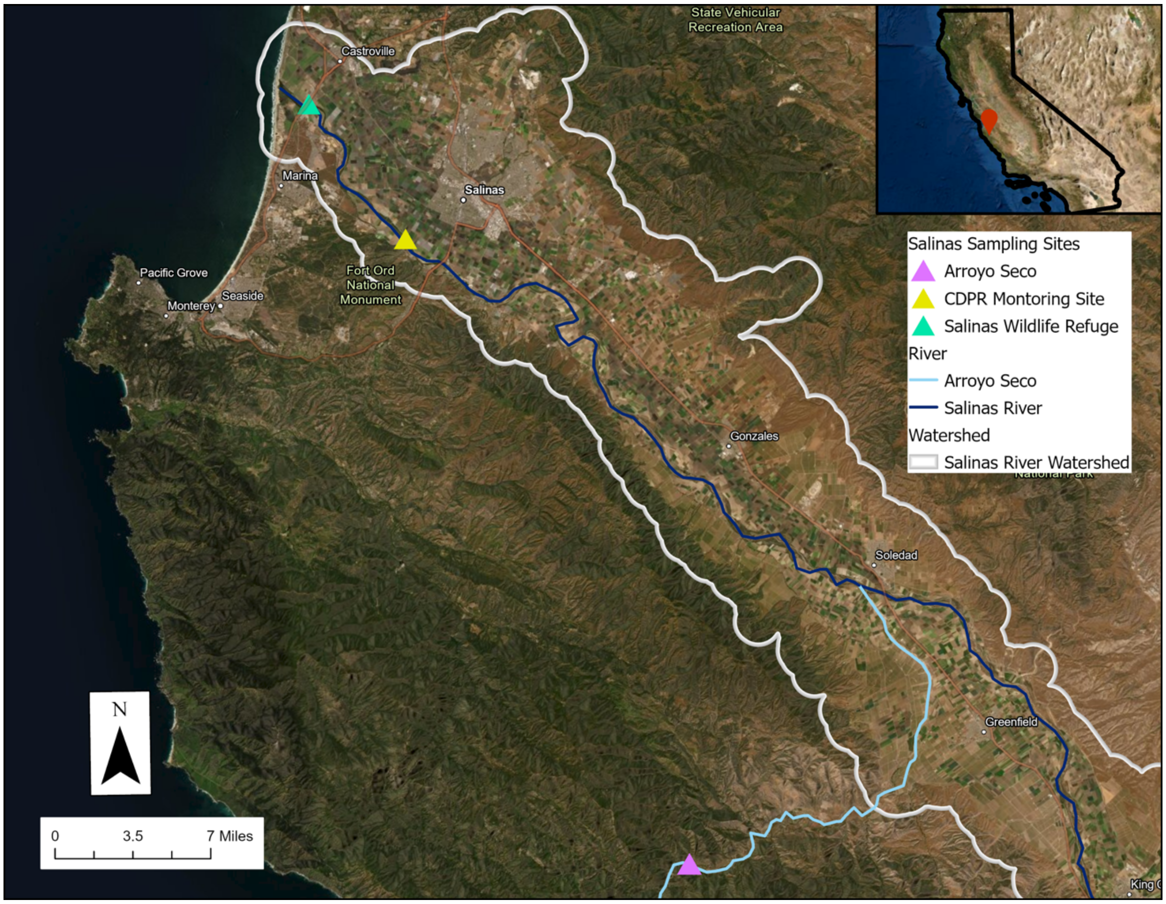

Figure 2. Map showing the three sampling sites within the Salinas River Valley (approximate delineation). A field bioreactor system in Salinas, CA on private agricultural land also served as a separate sampling location.

east from the Santa Lucia Range into the Salinas River (Taylor et al. 2017). Within this study, Arroyo Seco is categorized as a "comparison Salinas" sampling site.

Pesticide Remediating Bioreactor

The pesticide remediating bioreactor used within this study is located in Salinas, CA and has various unique features. Built in 2019 by John Silveus, its dimensions are 4 feet wide, 2 feet tall, and about 48 feet long (Figure 3). Within the bioreactor, sediment makes up the bottom portion of the trench, with a layer of woodchips as the mid-layer, and space at the top of the trench to allow water to pass through. The bioreactor has four sampling ports along its length, is completely above ground, and connected to a drainage system that collects runoff from nearby agricultural fields and pushes it through the

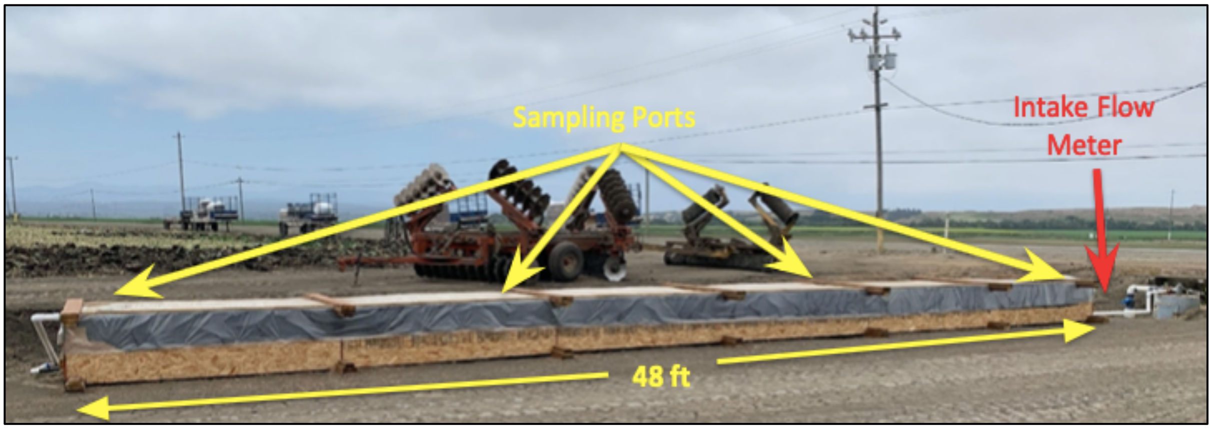

Figure 3. Picture of the bioreactor in Salinas, CA

bioreactor. The flow of the water runoff passing into the bioreactor is measured using an intake flow meter connected directly to the drainage system. In order to promote remediation, the bioreactor is dosed regularly with known bacterial remediators isolated from the environment (Mortensen et al. 2019).

Setup and Maintenance. To ensure proper maintenance before sampling, the bioreactor was flushed thoroughly to clean the system of any contaminants that may have collected over the past year. This was done by running water from the attached drainage system through the entire bioreactor for a whole day. This was done about 3 times before starting the sampling period to ensure the system was clean and ready for use. After the bioreactor was flushed it was dosed with multiple strains of pesticide-remediating bacteria that have been isolated, inoculated, and grown within a laboratory environment. This was completed several times throughout every sampling season. A comprehensive list of dosing dates and strains used can be found below (Table 1).

Table 1. Bioreactor dosing dates and strains used for growing seasons 2019-2021.

	Target Pesticide					
	Imidacloprid	Bifenthrin	Lambda Cyhalothrin	Malathion	Cypermethrin	Permethrin
Date Dosed	Bacteria Strain Used (2L)					
07-22-2019 07-29-2019	SV-BI-1W, SV-BI-1Y, SV-BI-2, SV-BI-3, SV-BI-5, SV-BI-6, SV-BI-7, SV-BI-8, SV-BI-9W, SV-BI-9Y					SV-BP-1, SV-BP-2, SV-BP-3, SV-BP-4, SV-BP-5, SV-BP- 7
8-14-2020 8-26-2020 09-16-2020	SV-BI-1W, SV-BI-3, SV-BI-5	SV-BB-2, SV-BB-6, SV-BB-9	SV-BLC-1, SV-BLC-2, SV-BLC-3, SV-BLC-4	SV-BM-1, SV-BM-4, SV-BM-8	SV-BC-2	SV-BP-1, SV-BP-5, SV-BP-7
7-02-2021	SV-BI-5b	SV-BB-2, SV-BB-9	SV-BLC-1, SV-BLC-4	SV-BM-1, SV-BM-4	SV-BC-6, SV-BC-8	SV-BP-5, SV-BP-7
7-10-2021 7-13-2021 7-23-2021 8-02-2021 8-08-2021 8-18-2021 9-29-2021	SV-BI-2a, SV-BI-5b	SV-BB-2, SV-BB-9	SV-BLC-1, SV-BLC-2, SV-BLC-3	SV-BM-1, SV-BM-4, SV-BM-8, SV-BM-10	SV-BC-6, SV-BC-8	SV-BP-5, SV-BP-7

Sample Collection

Salinas and Tijuana River Valleys Field Sampling

Sediment samples were collected from three sites near the Salinas River Valley and the Tijuana River Valley in January 2022 and February 2022, respectively. Each site had three corresponding sediment samples that were taken. All sites were sampled in an identical manner described below.

Upon arriving at a site, a YSI sonde was placed within the nearest body of water. Measurements were taken for turbidity, dissolved oxygen, salinity, temperature, and pH. Water was also tested using a HACH colorimeter to measure phosphate and nitrate concentrations. After all environmental metadata was collected, a 5 inch x 5 inch patch of sediment near the body of water was chosen to collect from, away from debris or any human disturbance. The top millimeter layer of sediment was removed using a sterile, plastic scoop spatula. The remaining sediment was then collected into a Sterile Five-O™ 5mL MacroTubes® and immediately placed on dry ice until being transferred to a -80°C freezer.

Bioreactor Field Sampling

Sampling from the bioreactor took place from 2019-2021 within the summer months to ensure proper flow of water from the growing season to push through the bioreactor. A "sampling run" was done over the course of at least three consecutive days, where the bioreactor was sampled three times a day, at 9am, 1pm, and 5pm. The bioreactor was only sampled when the water running through the system was at a rate of a 10, 20, or 30 gallons/minute at the beginning of the day. At each sampling time, environmental metadata was first collected from each port of the bioreactor using a YSI sonde and a HACH colorimeter. Measurements collected included temperature, salinity, pH, dissolved oxygen, conductivity, phosphate, nitrate, and turbidity.

A sediment sample was taken from each port of the bioreactor at every sampling run using a Johnny Jolter Pro Toilet Plunger System and an attached PVC pipe (Figure

4). Sediment collected from the bioreactor was placed in a half-gallon bucket for ~10 minutes to allow any water to separate from the sediment and then decanted to remove excess water. The remaining sediment was collected in Sterile Five-O™ 5 mL MacroTubes® and immediately placed on ice until being transferred to a -80°C freezer.

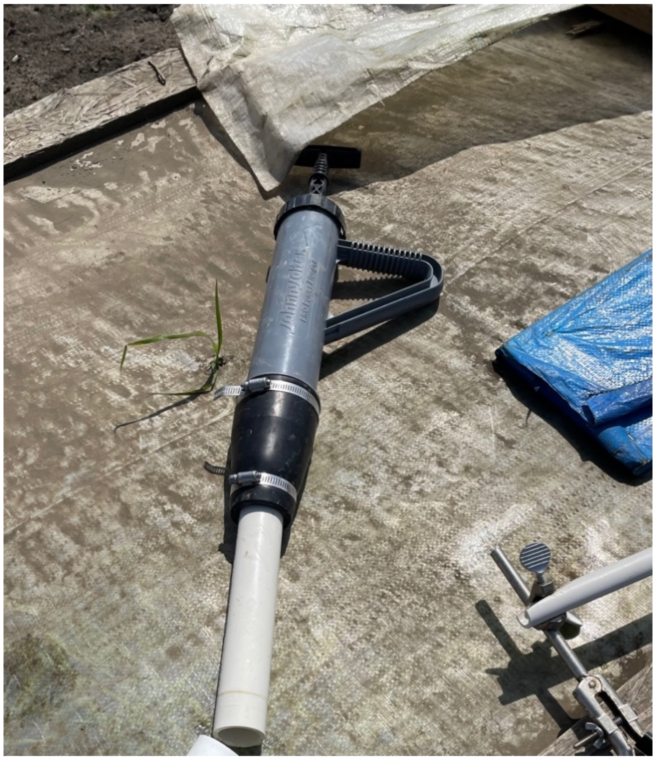

Figure 4. Picture of Johnny Jolter used to collect sediment samples from the bioreactor from 2020-2021.

Sample Processing

DNA from all sediment samples was extracted using the Qiagen DNeasy Power Soil Pro Kit. DNA extracts were quantified using a Thermo Scientific NanoDrop 2000C Spectrophotometer (Table A1) and gel electrophoresis. DNA extracts were processed using amplicon sequencing on the Illumina MiSeq System using the methods below. Upon quantifying all samples, 24 samples were chosen to be sent in to Novogene for whole metagenome shotgun sequencing (Table A2).

16S Amplicon Sequencing

Polymerase Chain Reactions. Polymerase chain reactions (PCRs) were performed on all selected DNA samples to amplify the 16S rRNA gene using the Thermo Scientific™ Phusion High-Fidelity PCR Master Mix with HF Buffer with a desired final concentration of 10-20 ng/μL for each 25 μL reaction. Each sample had a unique 319F and 806R dual-indexed PCR amplification primer added to the PCR as described by Fadrosh et al. 2014. Each sample had three replicate PCRs that were pooled together upon completion of the reaction. The following protocol was programmed on the thermocycler for every PCR:

Method: Calc

Lid: 105°C

Volume: 25 μL

1. 98°C, 30 seconds
2. 98°C, 15 seconds
3. 58°C, 15 seconds
4. 72°C, 15 seconds
5. Go to 2, 30X
6. 72°C, 1 minute
7. 4°C, ∞

PCR Clean-Up and Quantification. 65 microliters of every pooled PCR product were cleaned using the PCRClean DX purification system. The dual-sided magnetic bead clean-up followed the 96-well format method as described in Aline Biosciences Protocol Rev. 2.10.

After the dual-sided magnetic bead clean-up was completed for each PCR pooled product, the newly cleaned product's DNA concentration was quantified using the dsDNA HS Qubit Assay Kit. The qubit process followed the method as described by Invitrogen. The targeted concentration for the newly cleaned products was within 0.8 ng/μL – 10 ng/μL. If the concentration was above 10 ng/μL, the sample was diluted with molecular grade water to reach a concentration between 0.8 ng/μL – 10 ng/μL. If the concentration was below 0.8 ng/μL, all of the sample (26 μL) was included within the sample library.

Library Pooling. After each PCR product was quantified, a simple calculation was used to determine how much of each sample would be added to the final library pool. The target final concentration for each sample was 20 ng, which was divided by the qubit concentration to determine how many microliters of each sample was to be added to the pooled library. All samples with the desired volume were added to the same tube and served as the final library for sequencing.

Illumina MiSeq Library Preparation. The final library pool was quantified for its DNA concentration using Qubit. The library pool was concentrated using the CentriVap Benchtop Centrifugal Vacuum Concentrator at 37 °C for 24 minutes to match the desired concentration of 1.56 ng/μL based on the requirement of a 4nM library by Illumina. The concentrated library was denatured and diluted to an 8pM final concentration according to Protocol A: Standardized Normalization Method in the MiSeq Denature and Dilute Libraries Guide. A PhiX control was also denatured and diluted to a 20pM concentration according to the Denature and Dilute Phix Control in the MiSeq Denature and Dilute Libraries Guide. The final step of the library preparation involved combining the 8pM library and a 15% spike-in PhiX control together. The resulting solution was put into the V3 Illumina sequencing cartridge and placed into the MiSeq Illumina Sequencer.

Negative Controls. In order to account for contamination or index misassignment that is often seen within Illumina Next Gen Sequencing data, four negative controls were included within the sequencing run. Negative controls within this study involved amplifying 5uL of nuclease-free water using the same PCR settings as the experimental samples. The PCR products were then cleaned using dual-sided magnetic bead clean-up, quantified, and added to the final library pool along with the experimental samples.

Whole Metagenome Shotgun Sequencing

DNA extracted from 24 selected samples was sent in to NovoGene, a provider of genomic services, to be sequenced using shotgun metagenomic sequencing method. This sequencing technique provides data that can be used to identify functional profiling, predict genes, and determine microbial interactions within a microbial community. Each sample was sequenced to return a maximum of 6 Gb of data.

Data Analysis

16S Amplicon Data

Upon completion of amplicon sequencing using the Illumina MiSeq System, the returned raw fastq files were processed using Cutadapt (Martin 2011) to remove any primer and adapter sequences. The trimmed sequences were then processed using FastQC (Andrews 2010) to understand the quality of the sequences and inform the microbial analysis pipeline. QIIME2 (Boylen et al. 2019) a microbial analysis package, was used to filter and prepare sequences for diversity analyses and taxonomic assignments by returning ASVs (amplicon sequence variants) that represent different taxonomic classifications. Following similar methods followed by Nguyen et al. 2014, the number of reads of each ASV within the negative control sequenced samples was subtracted from the read abundances of the ASVs in the experimental samples.

Diversity Analyses. Alpha diversity of the processed ASVs was assessed using Observed and Shannon diversity measures within the phyloseq and vegan packages in R. Alpha diversity is a term that includes metrics used to understand the composition of an ecological community by quantifying the number of taxonomic groups and the way these groups are distributed within the environment (Willis 2019). Beta diversity of the processed ASVs was assessed using the Bray-Curtis Dissimilarity method (Chao et al. 2005) and three different ordination methods using the phyloseq package in R (McMurdie & Holmes 2011). Beta diversity is a term that includes metrics used to identify the differences, or similarities, between two or more groups by using a chosen distance metric to compare various diversity characteristics of the groups (Maziarz et al. 2018).

The methods used to evaluate beta diversity within this study were all non-parametric. Parametric statistical approaches within microbiology require various assumptions to be met by the biological dataset (Paliy & Shankar 2016). In order to meet necessary assumptions for parametric statistical approaches, original datasets often need to be transformed but may be negatively affected based on the type of variable analyzed: discrete or continuous (Paliy & Shankar 2016). Because non-parametric statistics are not associated with any sort of distribution, these analyses can be applied to non-transformed data (Paliy & Shankar 2016), which is the approach that was used in this study.

The first method used to assess beta diversity was a Principle Components Analysis (PCA). PCA uses Euclidean distances to visualize dissimilarity between samples based on the abundance of unique ASVs, which is greatly affected by a high number of zeros in a dataset and could display false distributions (Paliy & Shankar 2016). A second distance method, Bray-Curtis distance, was used to calculate dissimilarity among microbial communities from different sampling locations (Paliy & Shankar 2016). Bray-Curtis dissimilarity values were then ordinated using two ordination visualization methods: Principle coordinates analysis (PCoA) and non-metric multidimensional scaling (NMDS).

PCoA is similar to PCA by ordering objects along axes of principle components in an attempt to explain the variance in the original dataset (Paliy & Shankar 2016). Instead of using Euclidean distances, PCoA can be applied to any distance method. NMDS attempts to use a small number of ordination axes to fit the data to the chosen number of dimensions and the distances among all samples are ranked (Paliy & Shankar 2016). As long as the number of axes is below 4, all axes of variation will be displayed once the analysis is complete (Paliy & Shankar 2016). A stress parameter is then calculated to represent the fit between the observed distances and the resulting ordination.

Taxonomic Assignments. Upon completion of the QIIME2 microbial analyses pipeline, a unique primer-specific classifier was created using RESCRIPt (Robeson et al. 2021), a QIIME2 plug-in, that uses full-length SILVA 16s rRNA reference sequences to assign taxonomies. The classifier was then used against the processed ASVs.

Shotgun sequencing data

The raw fastq files returned from Novogene were analyzed for their quality using FastQC upon removing primers and adapter sequences using Cutadapt (Martin 2011). The following tools within the “Welcome to Metagenome Analysis 101” on KBase (Dow et al. 2021) were used to perform the following functions:

- metaSPAdes (Assembling reads into contigs) (Nurk et al. 2017)
- MetaBat2 (Binning contigs) (Kang et al. 2019)
- CheckM (Assessing genome quality) (Parks et al. 2014)
- DRep (Dereplicating genomes) (Olm et al. 2017)
- GTDB-Tk (Obtaining objective taxonomic assignments of genomes) (Chaumeil et al. 2019)
- DRAM (Annotating and distilling assemblies using the KEGG database) (Shaffer et al. 2020)

In addition to Kbase, OmicsBox (Biobam Bioinformatics 2019), a bioinformatics analysis platform, was used to process the annotated protein sequences from the resulting dereplicated high-quality dereplicated bins using the following built-in tools:

- Blast2GO (Götz et al. 2008)
- FatiGO (Al-Shahrour et al. 2004).

Gene Ontology (GO) elements were used within the above OmicsBox analyses. The GO is a way to describe genetic data using three aspects: molecular function, cellular components, and biological processes (Thomas 2016). A GO annotation includes possible gene function along with its presumed ontology function (Thomas 2016). GO annotations were assigned to each sequence within the dereplicated high-quality bins in OmicsBox.

After assigning GO annotations, grouping patterns were investigated using a hierarchical clustering calculation, within the heatmap.2 package in R (Warnes et al. 2015), of the relative abundance of each Level-3 GO category within each sample. Any evidence of grouping was tested with a two-tailed Fisher's Exact configuration (Al-Shahrour et al. 2004) to show significant differences between two groups.

Results

16s Amplicon Sequencing

Upon basecalling the returned paired-end amplicon sequences from the Illumina MiSeq system, individual fastq files were generated for both the forward and reverse reads of every sample. The forward and reverse adapter sequences were removed using Cutadapt and the resulting sequences were run through QIIME2 Demux Summary to produce a demultiplexed summary table (Table 2). Based on the quality plots (Figures 5 and 6) and their corresponding parametric seven-number summary tables (Tables 3 and 4) created in QIIME2, the forward and reverse reads were truncated at 258 bp and 170 bp, respectively, to ensure a median sequence quality score above 30. Upon truncation, filtering, merging, and removal of chimeric sequences (Table 5), 10,680 unique amplicon sequence variants (ASVs) were returned for all samples with varying lengths (Table 6).

Table 2. Demultiplexed sequence length summary.

Forward Reads		Reverse Reads	
Total Sequences Sampled	10000	**Total Sequences Sampled**	10000
2%	269 nts	**2%**	257 nts
9%	269 nts	**9%**	258 nts
25%	271 nts	**25%**	259 nts
50% (Median)	274 nts	**50% (Median)**	261 nts
75%	275 nts	**75%**	264 nts
91%	276 nts	**91%**	265 nts
98%	285 nts	**98%**	285 nts

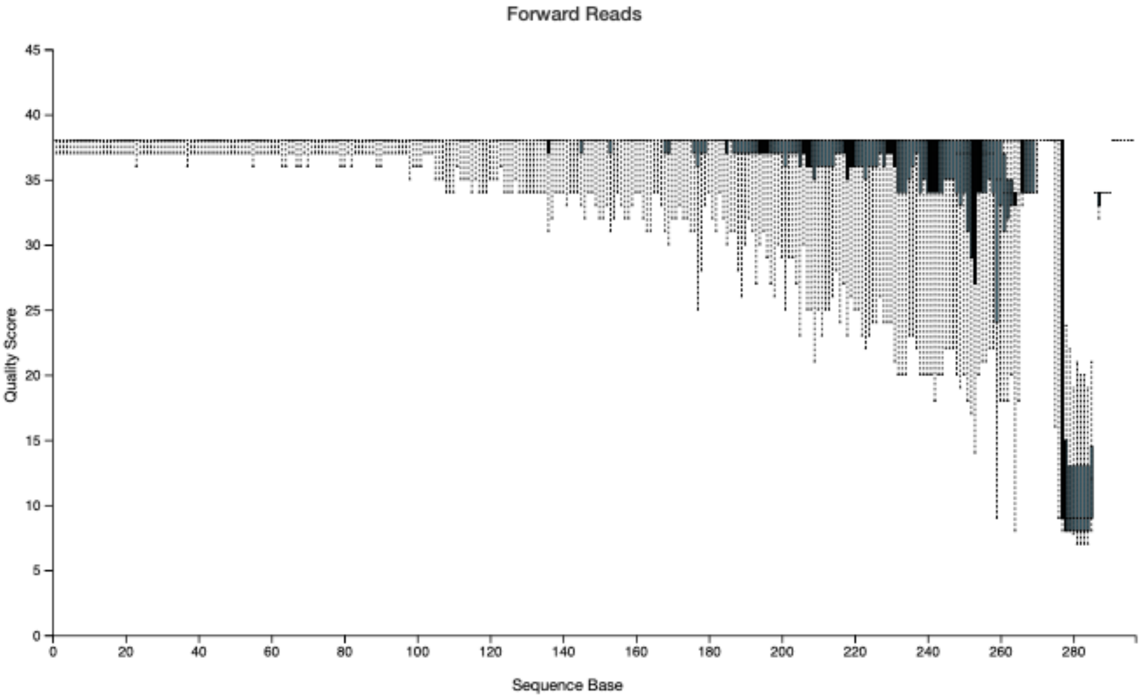

Figure 5. Box and whisker plot of sequence quality scores according to sequence base length of forward reads.

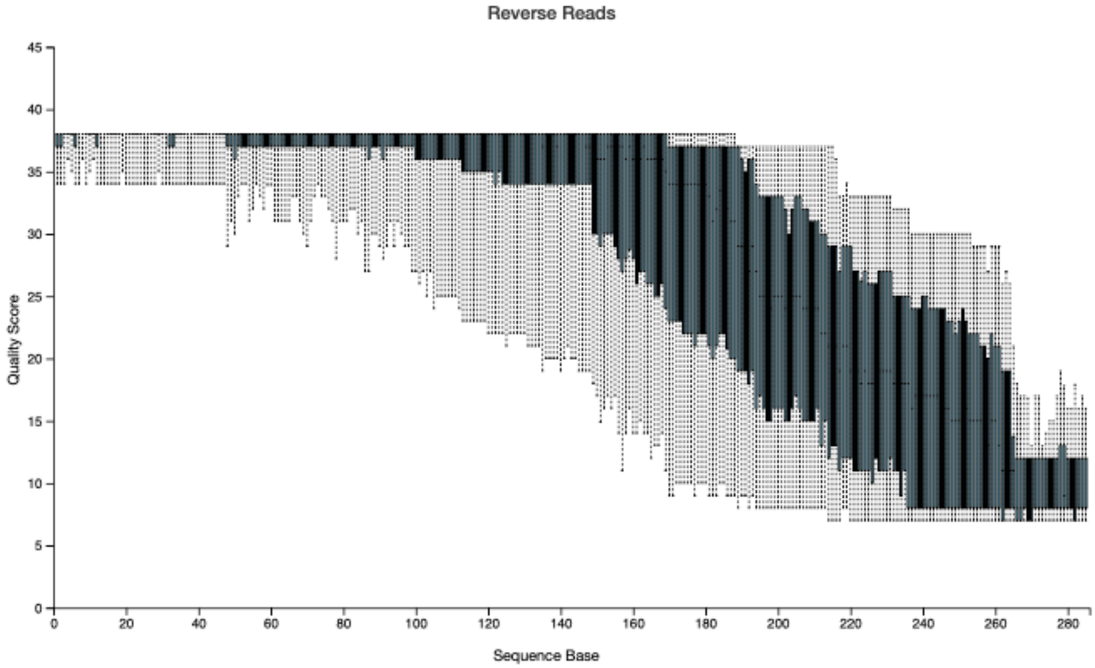

Figure 6. Box and whisker plot of sequence quality scores according to sequence base length of reverse reads.

Table 3. Parametric seven-number summary for forward reads at position 258.

Box plot feature	**Percentile**	**Quality score**
(Not shown in box plot)	2nd	8
Lower Whisker	9th	20
Bottom of Box	25th	33
Middle of Box	50th (Median)	35
Top of Box	75th	38
Upper Whisker	91st	38
(Not shown in box plot)	98th	38

Table 4. Parametric seven-number summary for position 170.

Box plot feature	**Percentile**	**Quality score**
(Not shown in box plot)	2nd	8
Lower Whisker	9th	9
Bottom of Box	25th	23
Middle of Box	50th (Median)	34
Top of Box	75th	37
Upper Whisker	91st	38
(Not shown in box plot)	98th	38

Table 5. Denoising statistics output generated in QIIME2.

Sample Name	Input	Filtered	Percentage of input passed filter	Denoised	Merged	Percentage of input merged	Non-chimeric	Percentage of input non-chimeric
AS01	34546	29641	85.8	26001	12241	35.43	10605	30.7
AS02	46179	40562	87.84	36955	19048	41.25	15959	34.56
AS03	36296	31518	86.84	28647	13501	37.2	10960	30.2
CONTROL1	9524	8	0.08	1	0	0	0	0
CONTROL2	11336	9741	85.93	7783	3412	30.1	2995	26.42
CONTROL3	26583	22106	83.16	21264	12084	45.46	11302	42.52
CONTROL4	12380	9957	80.43	9393	5003	40.41	4742	38.3
GC01	55496	49117	88.51	44773	12959	23.35	11661	21.01
GC02	10	5	50	1	0	0	0	0
GC03	52718	46534	88.27	43113	9768	18.53	8784	16.66
GC04	68579	61565	89.77	58306	15262	22.25	13184	19.22
GC05	23873	19458	81.51	17239	8342	34.94	7542	31.59
GC06	20141	15728	78.09	13624	6835	33.94	6234	30.95
PL01	27225	23383	85.89	20538	9698	35.62	8709	31.99
PL02	41602	35414	85.13	30185	9387	22.56	7974	19.17
PL03	43313	37560	86.72	32309	10519	24.29	8918	20.59
SB148	53717	47305	88.06	44978	21963	40.89	17720	32.99
SB160	52077	45761	87.87	43374	21065	40.45	16636	31.95
SB456	53552	47072	87.9	44735	13275	24.79	12158	22.7
SB460	42276	36659	86.71	34364	10340	24.46	9224	21.82
SB508	67391	60213	89.35	56562	25844	38.35	22156	32.88
SB516	57120	51154	89.56	48078	21806	38.18	19882	34.81
SB596	59978	53091	88.52	49963	17812	29.7	16315	27.2
SB604	42312	37104	87.69	34193	12583	29.74	11513	27.21
SB666	8603	5952	69.19	4558	2100	24.41	1793	20.84
SB670	31363	27460	87.56	24943	12845	40.96	11633	37.09
SB931	43016	21914	50.94	19990	9938	23.1	9206	21.4
SB935	5	2	40	2	0	0	0	0
SG01	37514	33212	88.53	29696	10624	28.32	9294	24.77
SG02	36882	32689	88.63	29394	11835	32.09	10571	28.66
SG03	57136	50386	88.19	46634	19019	33.29	16925	29.62
SRA01	46203	38518	83.37	34291	9979	21.6	8937	19.34
SRA02	80711	70521	87.37	65886	13592	16.84	12425	15.39
SRA03	22703	19442	85.64	17162	3852	16.97	3453	15.21
SRB01	53339	46137	86.5	40266	13928	26.11	11454	21.47
SRB02	57053	48645	85.26	43153	15754	27.61	12878	22.57
SRB03	54710	47555	86.92	42957	18622	34.04	15291	27.95

Table 6. Representative Sequences Truncation Output

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
10680	258	416	404.73	158	11.99

Negative Controls

Overall, 13% of the total filtered and denoised ASVs were found within negative control samples (Table 7). This resulted in a loss of 1,168 low abundance ASVs (Table 7). Upon removing the low abundance ASVs, the negative control samples were also removed from the dataset before rarefaction.

Table 7. Negative control ASVs summary.

Total unique ASVs	Number of ASVs found within Negative Controls	Percentage of ASVs found within Negative Controls	Unique ASVs Removed with 0 reads
10680	1476	13	1168

Rarefaction

All experimental samples were rarefied at an even sampling depth of 3,000 reads based on the resulting rarefaction curve (Figure 7). Rarefaction is a method used to adjust for differences in library sizes between samples (Willis 2019). Rarefaction resulted in a loss of 3 samples, GC02, SB666, SB935, due to total read numbers being below the 3,000 read sampling depth. Rarefied samples were assessed for alpha diversity, beta diversity, and taxonomic classifications.

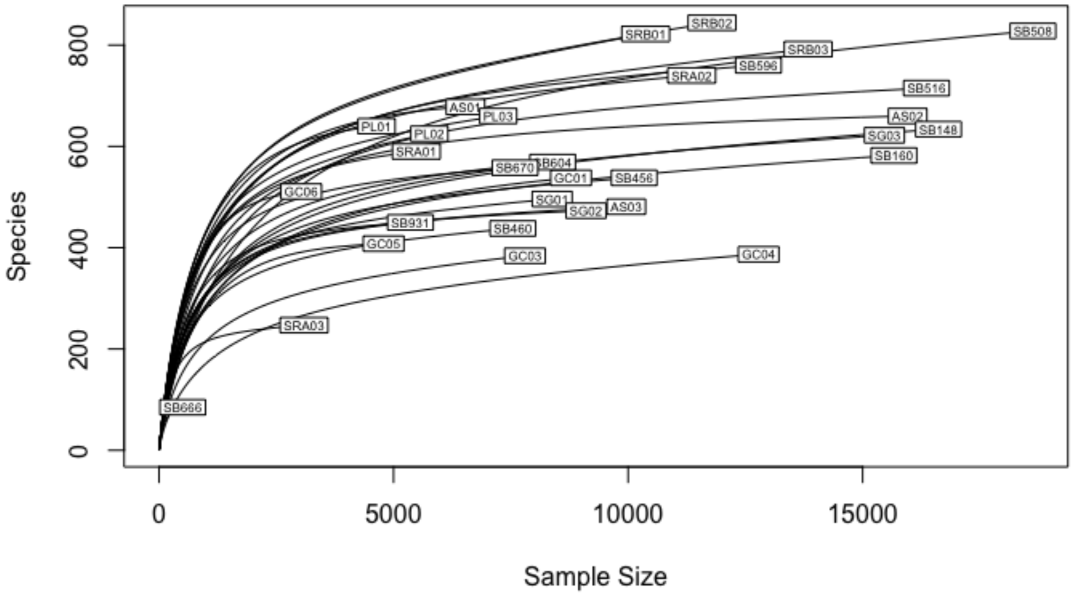

Figure 7. Rarefaction curve of experimental sample total reads and observed species richness.

Alpha Diversity

Location and Comparison Category. Observed and Shannon diversity measure values were highest within the CDPR Salinas Site and lowest within Goat Canyon (Figure 8 and Table 9). After grouping samples based on comparison category (Table 8), Observed and Shannon diversity values were highest within the polluted Salinas group and comparison Tijuana group, respectively (Figure 9 and Table 10). Observed and Shannon diversity values were lowest within the polluted Tijuana group (Table 9). Results showed a significant difference in Observed and Shannon diversity values within all location sites (Tables 11 and 12; $p<0.05$). Results also showed a significant difference in Observed and Shannon diversity values within comparison categories (Tables 13 and 14; $p<0.05$).

Table 8. Comparison categories based on sample ID prefix.

Sample ID	**Location Site**	**Comparison Category**
AS	Arroyo Seco	Comparison Salinas
GC	Goat Canyon	Polluted Tijuana
PL	Penasquitos Reserve	Comparison Tijuana
SB	Bioreactor	Bioreactor
SG	Smuggler's Gulch	Polluted Tijuana
SRA	Salinas River Wildlife Refuge	Polluted Salinas
SRB	Salinas CDPR Site	Polluted Salinas

Table 9. Summary table median alpha diversity by location.

	Alpha Diversity	
Location Site	**Observed**	**Shannon**
Arroyo Seco	546	5.38
Bioreactor19	431	5.33
Bioreactor20	518	5.14
Bioreactor21	460	5.62
CDPR Salinas Site	638	6.01
Goat Canyon	391	4.58
Penasquitos Reserve	575	6
Salinas River Refuge	551	5.92
Smugglers Gulch	419	5.3

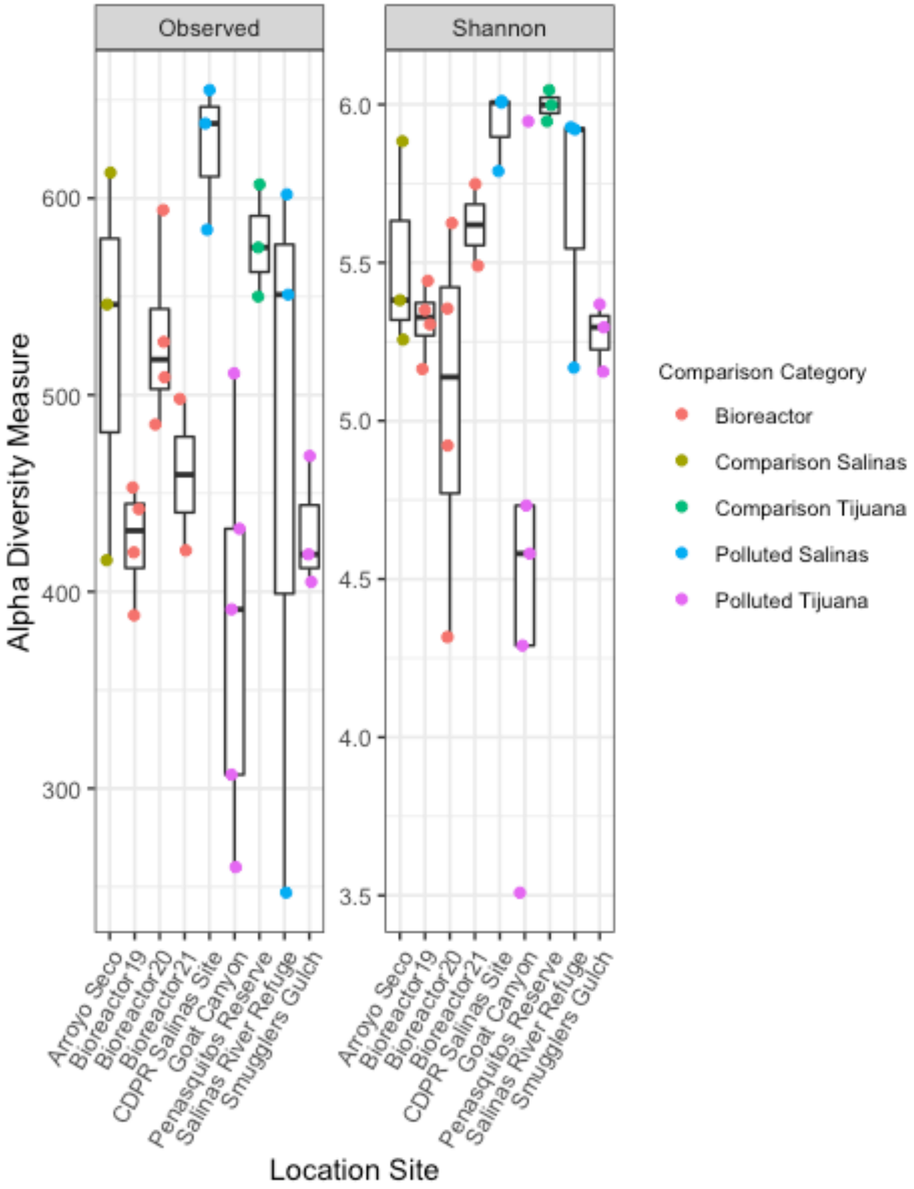

Figure 8. Alpha diversity measures for samples based on location site and color-coded by comparison category.

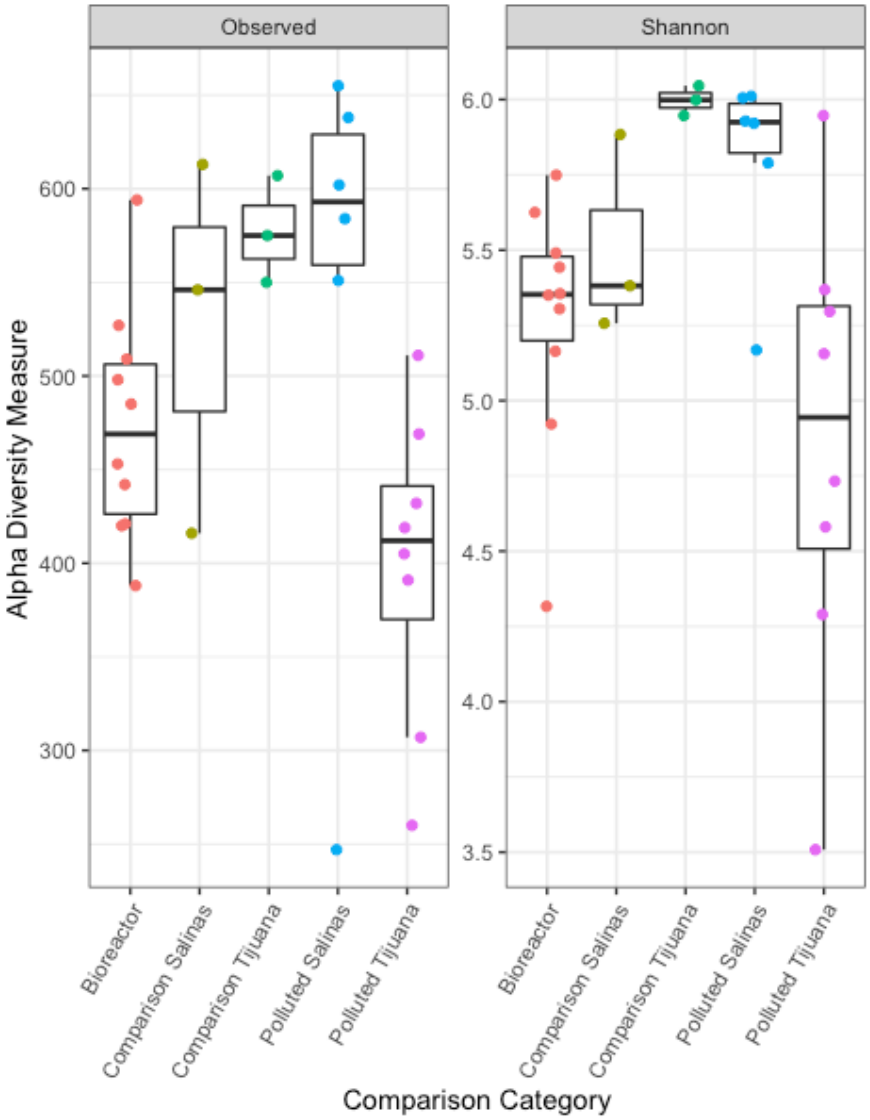

Figure 9. Alpha diversity measures for samples based on comparison category.

Table 10. Summary table of median alpha diversity measure values by comparison category.

	Alpha Diversity	
Comparison Category	**Observed**	**Shannon**
Bioreactor	469	5.35
Comparison Salinas	546	5.38
Comparison Tijuana	575	6
Polluted Salinas	593	5.92
Polluted Tijuana	412	4.94

Table 11. One-way ANOVA results: Observed diversity values of location site.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Location Site	8	177336	22167	3.409	0.0115
Residuals	21	136566	6503		

Table 12. One-way ANOVA results: Shannon diversity values of location site.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Location Site	8	6.122	0.7653	3.295	0.0135
Residuals	21	4.877	0.2322		

Table 13. One-way ANOVA results: Observed diversity values of comparison category.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Comparison Category	4	103731	25933	3.085	0.0341
Residuals	25	210171	8407		

Table 14. One-way ANOVA results: Shannon diversity values of comparison category.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Comparison Category	4	4.633	1.1584	4.549	0.00673
Residuals	25	6.366	0.2546		

Environmental Metadata. Environmental metadata collected during sampling were analyzed through an Analysis of Covariance (ANCOVA) to measure the effect of interactions between categorical and continuous variables. Environmental metadata collected included temperature, dissolved oxygen, pH, phosphate and nitrate concentrations, turbidity, and salinity which are all continuous variables. The first assumption that needs to be met within the dataset is evidence of an independent relationship between the covariates (environmental metadata) and the treatment (Location Site). This assumption was tested using an ANOVA model where every environmental measurement variable was analyzed with the Location Site variable. All environmental variables had a p-value of < 0.05 indicating that they were not independent from the Location Site variable (Table A4). Because the environmental metadata was not proven to be independent, continuing with an ANCOVA analysis would produce inaccurate results when attempting to control the covariate, environmental metadata, within an ANCOVA model that analyzes the response variable (Alpha Diversity) and the treatment (Location Site) (Miller & Chapman 2001).

Beta Diversity

Principle Components Analysis. Several ordination and distance methods were used to visualize beta diversity within all location sites and comparison category groups. A principle components analysis (PCA) was used to visualize the dissimilarity between samples taken at different location sites based on unique ASV abundances (Figure 10). PC1 and PC2 together explained 75% of the variance (Table 15, Figure 11).

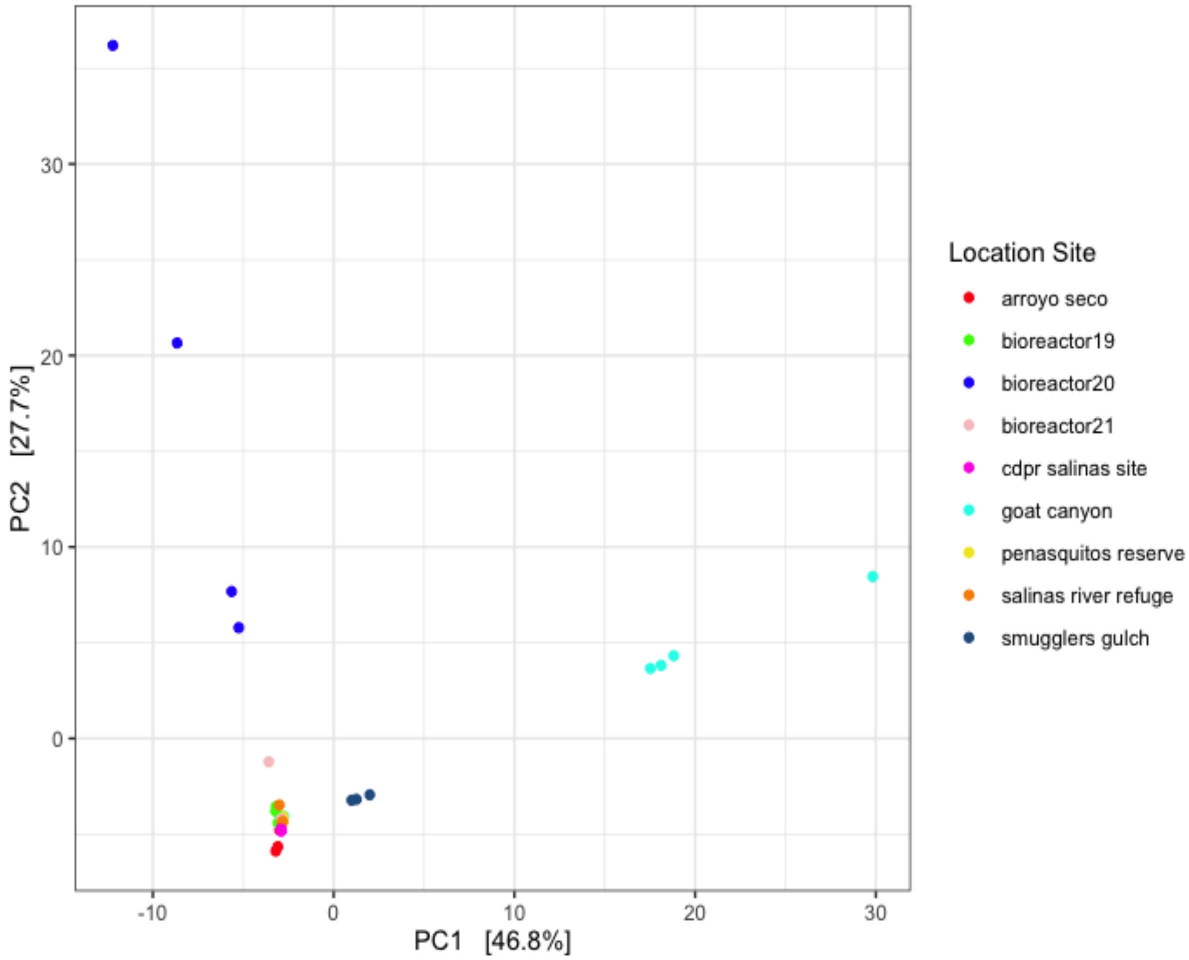

Figure 10. Beta diversity of samples based on composition of unique ASVs using Principal Components Analysis (PCA) method.

Table 15. Resulting principle component importance values.

	Importance PC1	**Importance PC2**	**Importance PC3**
Eigen value	88404	52196	10724
Proportion Explained	0.47	0.28	0.06
Cumulative Proportion	0.47	0.74	0.8

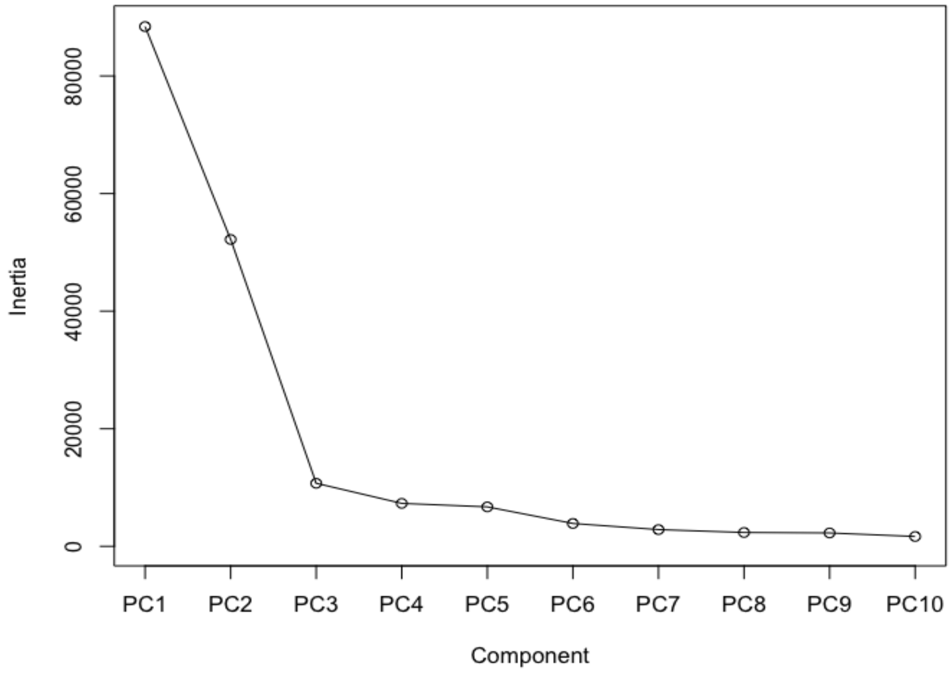

Figure 11. Resulting scree plot showing importance of PC1-PC10.

Location and Comparison Category. The Bray-Curtis Dissimilarity distance method was used to calculate dissimilarity between samples, and represent beta diversity, based on overabundant counts of ASVs within samples from different location sites. Bray-Curtis dissimilarity values were then input into a PERMANOVA to understand how different environmental factors affect beta diversity. Results showed a significant difference in beta diversity within all location sites and comparison categories (Tables 16, 17; p<0.05).

Table 16. PERMANOVA results: Beta diversity values of location site.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Location Site	8	7.587	0.64324	4.7328	1.00E-04
Residual	21	4.208	0.35676		
Total	29	11.795	1		

Table 17. PERMANOVA results: Beta diversity values of comparison category.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Comparison Category	4	4.3053	0.36501	3.5927	1.00E-04
Residual	25	7.4897	0.63499		
Total	29	11.795	1		

Environmental Metadata. All environmental metadata collected during sampling was also put through a PERMANOVA to understand any relationships to beta diversity. Nitrate concentrations, phosphate concentrations, dissolved oxygen, pH, temperature, and turbidity had a significant effect on beta diversity values (Table 18; $p<0.05$).

Table 18. PERMANOVA results: Beta diversity values of environmental metadata

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Nitrate	1	0.7431	0.063	1.8825	0.0075
Residual	28	11.052	0.937		
Total	29	11.795	1		
Phosphate	1	1.2448	0.10553	3.3035	1.00E-04
Residual	28	10.5503	0.89447		
Total	29	11.795	1		
Dissolved Oxygen	1	1.0205	0.08652	2.6519	0.0013
Residual	28	10.7746	0.91348		
Total	29	11.795	1		
pH	1	1.3343	0.11312	3.5715	1.00E-04
Residual	28	10.4607	0.88688		
Total	29	11.795	1		
Temperature	1	1.2987	0.11011	3.4645	1.00E-04
Residual	28	10.4963	0.88989		
Total	29	11.795	1		
Turbidity	1	0.9705	0.08228	2.5104	4.00E-04
Residual	28	10.8246	0.91772		
Total	29	11.795	1		

Non-Metric Multidimensional Scaling. A non-metric multidimensional scaling (NMDS) ordination visualization method was used to show the dissimilarity between samples based on unique ASVs. An NMDS stress plot was first created to determine how many dimensions would be appropriate for the ordination (Figure 12). The resulting NMDS ordination had 3 dimensions and a stress value of 0.08 (Figure 13). The observed dissimilarity was also plotted against the ordination distance to observe the fit of the value (Figure 14). The R^2 value of the correlation between the ordination values and predicted regression line ordination values line was 0.937 (Figure 14).

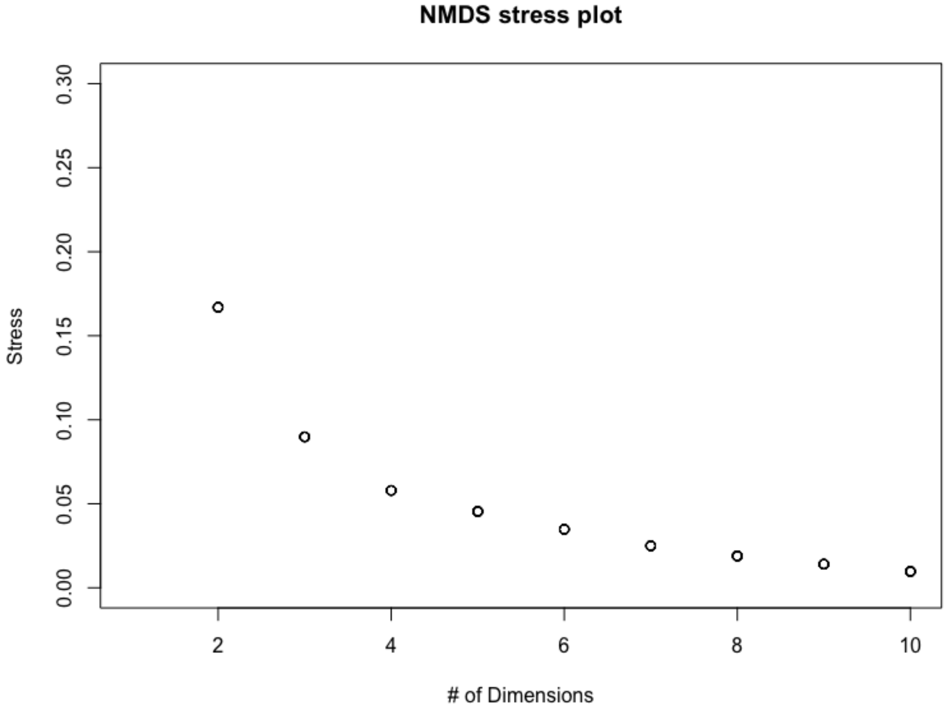

Figure 12. Stress related to the number of dimensions chosen for every unique trial NMDS ordination.

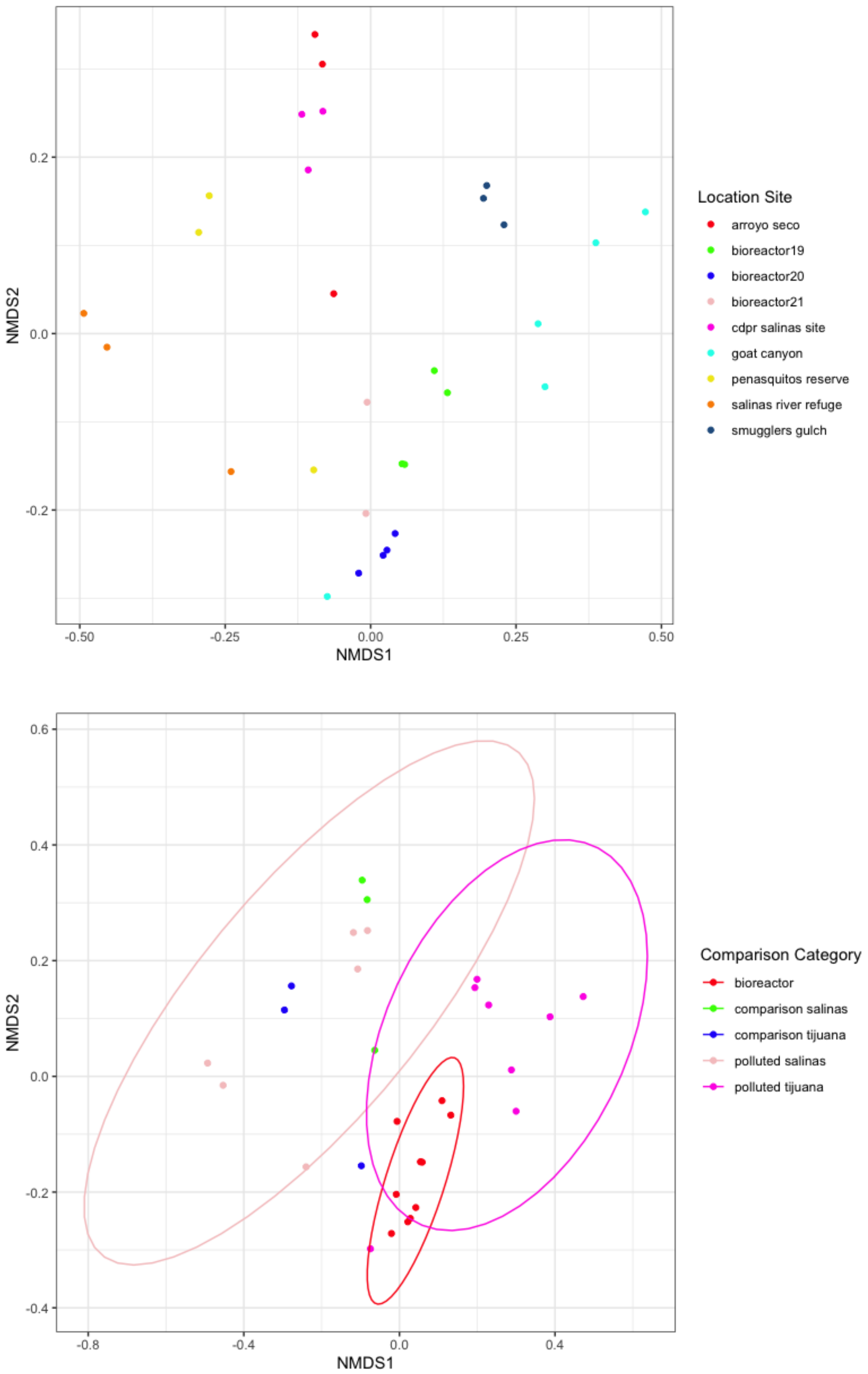

Figure 13. Beta diversity of samples based on ASVs using Bray-Curtis dissimilarity distance method and the first two dimensions of the NMDS ordination for location site and comparison categories (Dimensions: 3, Stress: 0.08983996). Ellipses produced on the plot have a 95% confidence interval

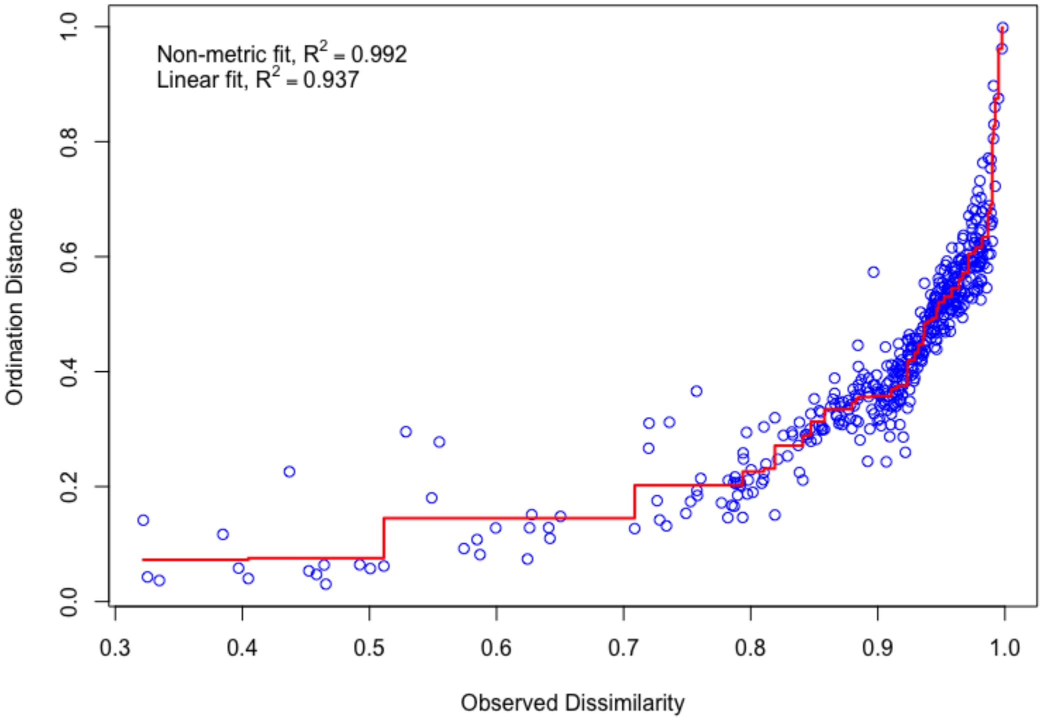

Figure 14. Stress plot of observed dissimilarity and NMDS ordination method.

Principal Coordinates Analysis. The second ordination visualization method used to show the Bray-Curtis Dissimilarity distance values was Principal Coordinates Analysis (PCoA). PCoA was used to visualize the dissimilarity of samples taken at different location sites (Figure 15). Axis1 and Axis2 together explained 28% of the variance (Table 19, Figure 15).

Table 19. Resulting principle coordinates axes importance values.

	Importance Axis 1	**Importance Axis 2**	**Importance Axis 3**	**Importance Axis 4**	**Importance Axis 5**
Eigenvalue	1.78	1.593	1.239	0.984	0.825
Relative Eig	0.1509	0.1351	0.105	0.0834	0.07
Cumulative Eig	0.151	0.286	0.391	0.474	0.544

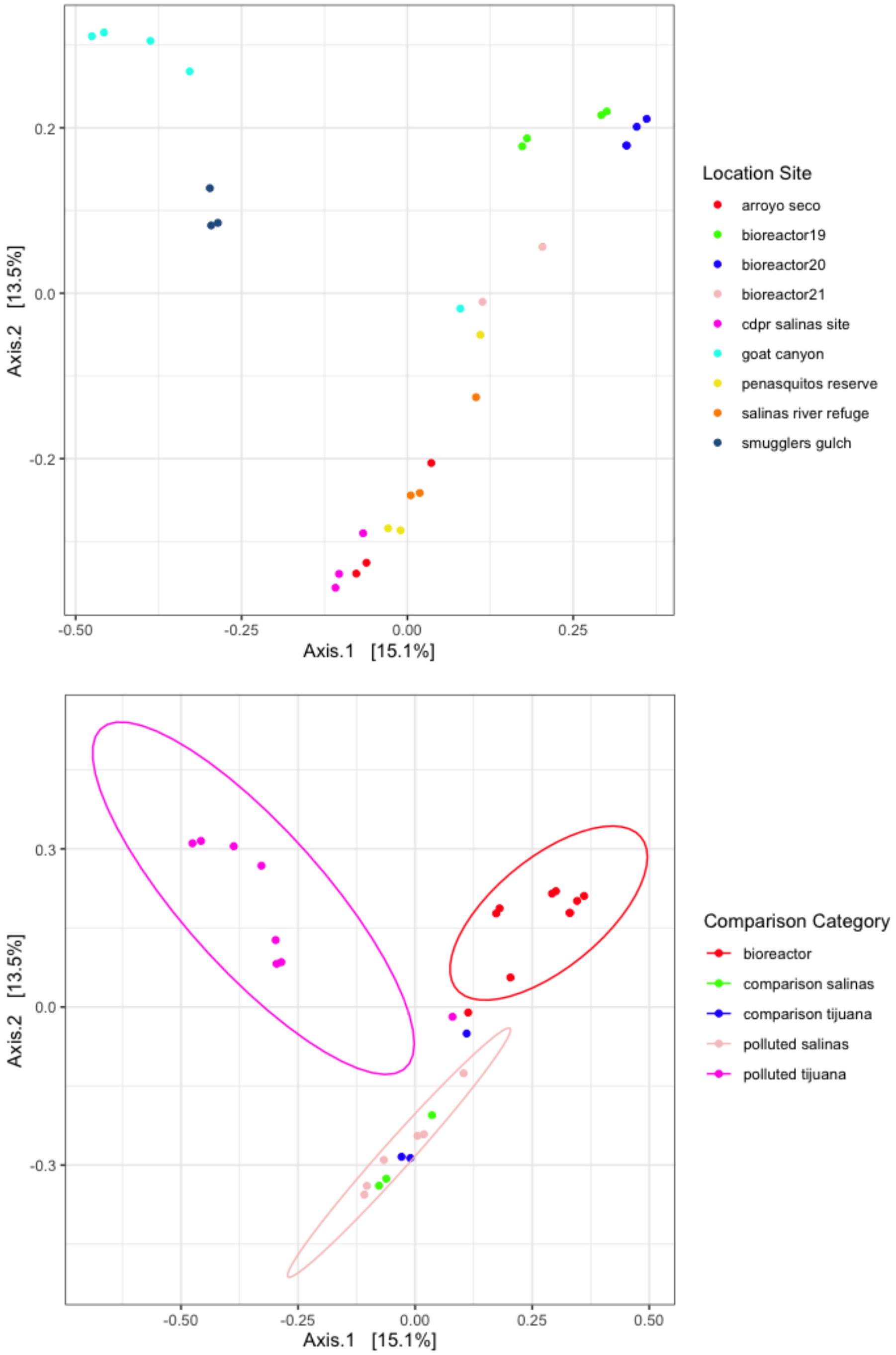

Figure 15. Beta diversity of samples based on ASVs using Bray-Curtis dissimilarity distance method and PCoA ordination. Ellipses produced on the plot have a 95% confidence interval

Taxonomic Classifications

A primer specific classifier was built and trained within QIIME2 using the Silva 138 99% OTUs full-length sequences with a confidence level of 70% and above. This primer specific classifier was used to assign taxonomy to each unique ASV. After the top 20 taxonomic assignments were subset from the rarefied samples, 5 Level 2 (phylum) taxonomic assignments and 9 Level 9 (Family-Genus) taxonomic assignments were displayed (Figure 16, Figure 17).

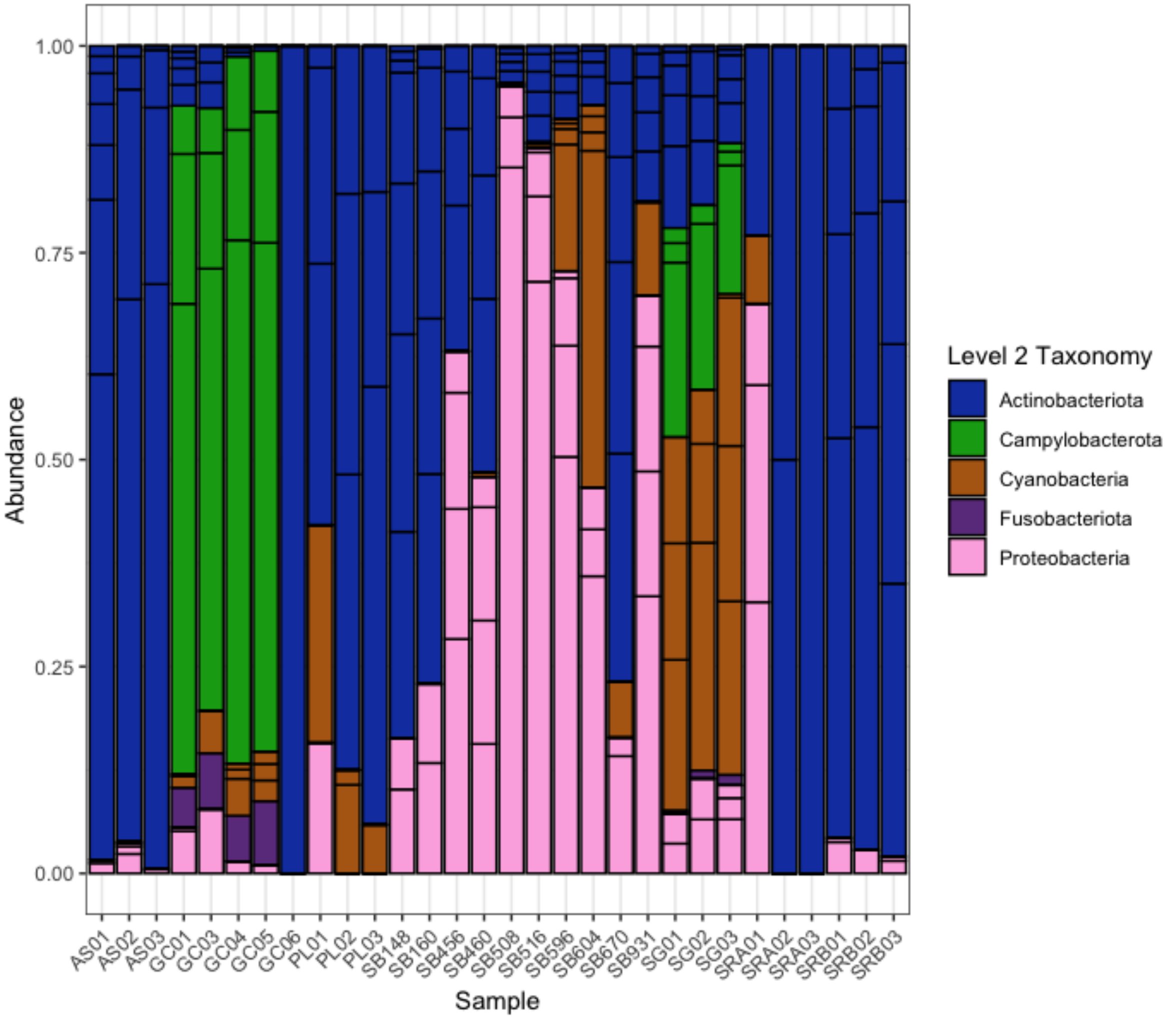

Figure 16. Top level 2 (Phylum) relative abundance taxonomic classifications of samples. Total ASVs (7418), total assigned at Level 2 (7365), 53 unclassified at level 2 (99.2% assigned).

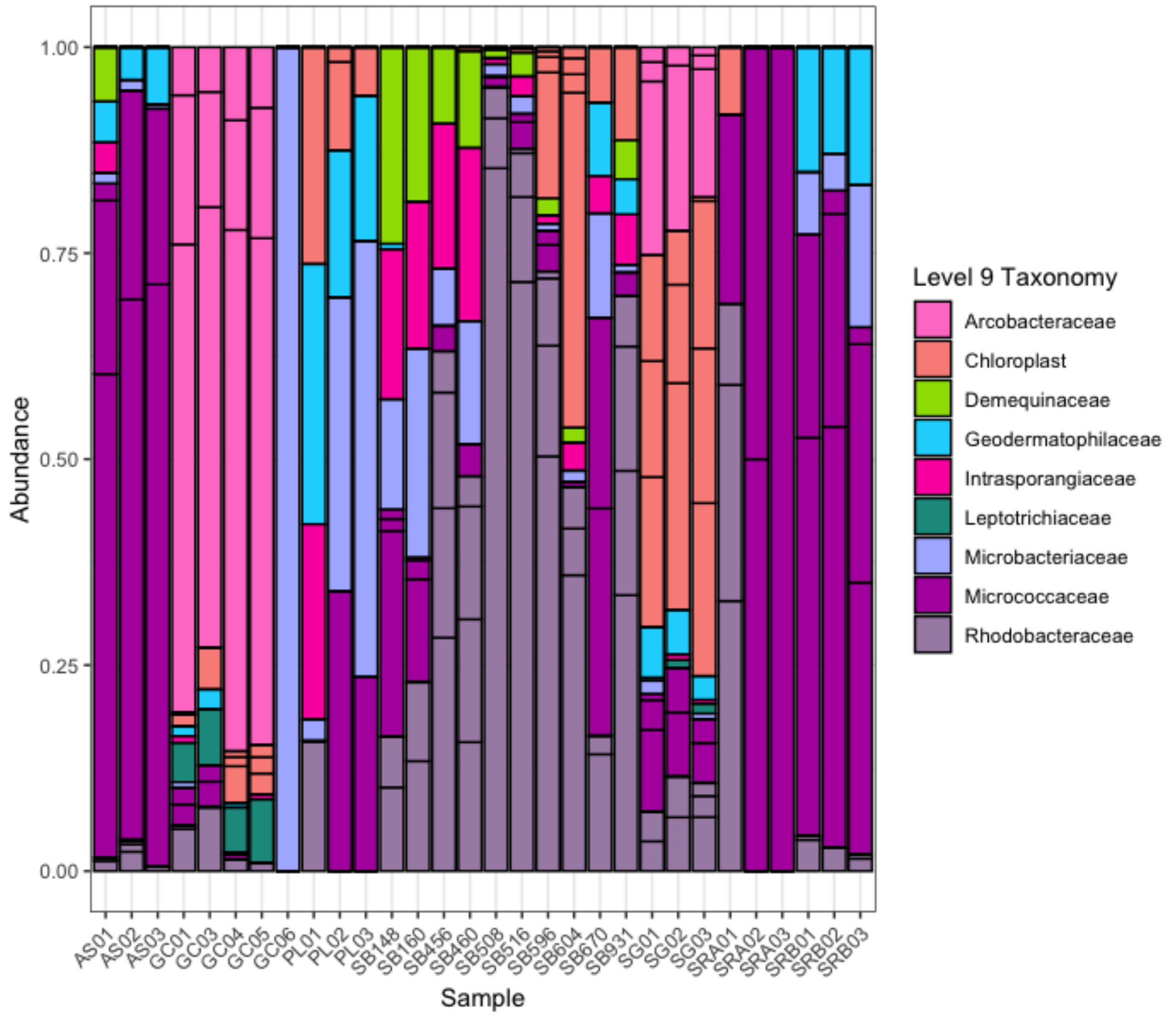

Figure 17. Top level 9 (Family-Genus) relative abundance taxonomic classifications of samples. Total ASVs (7418), total assigned at Level 9 (7066), 352 unclassified at level 9 (95.2% assigned).

Whole Metagenome Shotgun Sequencing

Upon receiving the shotgun sequencing data from Novogene, the quality scores of the samples were assessed using a table provided by Novogene (Table 20). All samples were selected for downstream analyses based on their low error percent and Q30 percent being within 90-100 ensuring high quality input sequences (Table 20). The forward and reverse adapter sequences were then removed using Cutadapt and the resulting sequences were input into various tools within KBase for downstream analyses.

Table 20. Summary table of sequencing data information provided by Novogene.

Sample	Location Site	Raw reads	Raw data	Effective(%)	Error(%)	Q20(%)	Q30(%)	GC(%)
GC4	Goat Canyon	42228580	6.3	99.74	0.03	97.9	93.82	50.99
GC6	Goat Canyon	53895040	8.1	99.76	0.02	98.14	94.42	50.27
PL2	Penasquitos Reserve	68175136	10.2	99.8	0.02	97.95	94.35	62.82
PL3	Penasquitos Reserve	79882038	12	99.81	0.02	98.05	94.65	63.21
SB148	Bioreactor19	64512674	9.7	99.82	0.02	97.86	94.45	63.52
SB160	Bioreactor19	70337882	10.6	99.85	0.03	97.77	94.26	64.34
SB456	Bioreactor19	64900930	9.7	99.79	0.03	97.79	94.07	58.85
SB460	Bioreactor19	56169934	8.4	99.77	0.02	97.99	94.64	59.33
SB508	Bioreactor20	64218388	9.6	99.8	0.02	97.95	94.36	61.5
SB516	Bioreactor20	80079828	12	99.84	0.03	97.83	94.3	60.91
SB596	Bioreactor20	67115462	10.1	99.79	0.03	97.75	93.79	60.83
SB604	Bioreactor20	60798524	9.1	99.81	0.03	97.69	93.93	60.42
SB666	Bioreactor21	72806202	10.9	99.82	0.03	97.85	94.18	64.31
SB670	Bioreactor21	69265510	10.4	99.8	0.03	97.81	93.97	63.36
SB931	Bioreactor21	65913716	9.9	99.8	0.03	97.79	94.28	61.53
SB935	Bioreactor21	74896116	11.2	99.71	0.02	98.01	94.51	61.24
SG2	Smugglers Gulch	92284358	13.8	99.81	0.02	97.88	94.53	62.65
SG3	Smugglers Gulch	50652200	7.6	99.73	0.03	97.89	94.2	60.92
SRA1	Salinas River Refuge	58769838	8.8	99.75	0.02	97.99	94.3	56.86
SRA3	Salinas River Refuge	52303720	7.8	99.73	0.03	97.83	93.93	58.57
SRB2	CDPR Salinas Site	67211948	10.1	99.74	0.03	97.83	94.08	63.15
SRB3	CDPR Salinas Site	57617388	8.6	99.76	0.02	97.95	94.35	63.14

Assembling Genomes, Filtering, and Dereplication

Resulting sequences were input into the metaSPAdes tool in KBase to assemble raw reads into contigs. The resulting number of contigs for every sample can be seen in Table 21. After reads were assembled, the MetaBAT2 tool in KBase was used to bin the newly created contigs based on sequence similarity into metagenomically assembled genomes (MAGs). The resulting MAGs were filtered based on completeness and contamination and only MAGs with a quality score of 40 or above were used within further downstream analyses as were labeled as “high quality” (Table 22). The resulting high-quality MAGs (n=217) were then input into the dRep tool in KBase to ensure downstream analyses of unique dereplicated MAGs. This resulted in the 217 high-quality MAGs being divided into 60 clusters with one unique MAG representing the cluster based on completeness and contamination. The 60 high-quality dereplicated MAGs had a number of sequences ranging from 28 to 1,497 with the MAGs from the bioreactor and the Salinas River Wildlife Refuge being the top 10 in sequence number and summation length of sequences (Table 23).

Table 21. Summary table of resulting contigs from metaSPAdes.

Sample	**Location Site**	**Total Contigs**
SRB02	CDPR Salinas Site	1066
SRB03	CDPR Salinas Site	3605
PL02	Penasquitos Reserve	4103
PL03	Penasquitos Reserve	6582
SG03	Smugglers Gulch	8137
SB670	Bioreactor21	11117
SG02	Smugglers Gulch	14047
SB666	Bioreactor21	14362
GC04	Goat Canyon	15794
SRA03	Salinas River Refuge	19375
SB931	Bioreactor21	21150
GC06	Goat Canyon	23243
SB604	Bioreactor20	23262
SB935	Bioreactor21	23526
SB596	Bioreactor20	24505
SB508	Bioreactor20	26303
SRA01	Salinas River Refuge	26521
SB148	Bioreactor19	28115
SB460	Bioreactor19	31784
SB516	Bioreactor20	33322
SB456	Bioreactor19	38487
SB160	Bioreactor19	38541

Table 22. Summary table of resulting bins from MetaBAT2.

Sample	**Location Site**	**Total MAGs**	**High Quality MAGs**
GC04	Goat Canyon	25	11
GC06	Goat Canyon	35	0
PL02	Penasquitos Reserve	3	0
PL03	Penasquitos Reserve	37	15
SB148	Bioreactor	33	7
SB160	Bioreactor	40	15
SB456	Bioreactor	55	27
SB460	Bioreactor	49	23
SB508	Bioreactor	36	21
SB516	Bioreactor	49	20
SB596	Bioreactor	40	15
SB604	Bioreactor	35	14
SB666	Bioreactor	20	3
SB670	Bioreactor	17	6
SB931	Bioreactor	26	9
SB935	Bioreactor	26	5
SG02	Smugglers Gulch	21	6
SG03	Smugglers Gulch	15	5
SRA01	Salinas River Refuge	31	8
SRA03	Salinas River Refuge	27	6
SRB02	CDPR Salinas Site	2	0
SRB03	CDPR Salinas Site	3	1

Table 23. List of resulting high quality dereplicated MAGs.

MAG	Location Site	Number of Sequences	Sum Length	Min Length	Average Length	Max Length
Concatenated MAGs	All MAGs	20,813	213,613,447	2,500	10,263.50	621,844
GC4bin11	Goat Canyon	247	2,102,587	2,518	8,512.50	38,999
GC4bin12	Goat Canyon	195	3,056,468	2,512	15,674.20	91,785
GC4bin17	Goat Canyon	142	2,566,124	2,564	18,071.30	94,349
GC4bin23	Goat Canyon	169	1,336,956	2,503	7,911	46,783
GC4bin3	Goat Canyon	236	1,781,117	2,528	7,547.10	42,751
GC4bin6	Goat Canyon	272	2,756,042	2,506	10,132.50	59,165
Pl3bin3	Penasquitos Reserve	362	2,708,654	2,512	7,482.50	42,765
Pl3bin35	Penasquitos Reserve	308	3,346,823	2,502	10,866.30	92,556
SB148bin15	Bioreactor19	217	2,262,142	2,504	10,424.60	54,491
SB148bin25	Bioreactor19	59	3,375,025	2,609	57,203.80	518,880
SB160bin11	Bioreactor19	485	2,623,361	2,501	5,409	23,901
SB160bin2	Bioreactor19	361	2,705,350	2,512	7,494	42,765
SB160bin3	Bioreactor19	224	2,188,170	2,518	9,768.60	73,258
SB160bin32	Bioreactor19	234	2,881,647	2,542	12,314.70	47,209
SB160bin40	Bioreactor19	238	1,693,953	2,637	7,117.40	24,497
SB160bin6	Bioreactor19	218	3,118,513	2,534	14,305.10	57,774
SB456bin11	Bioreactor19	602	5,826,228	2,500	9,678.10	51,527
SB456bin13	Bioreactor19	396	3,951,570	2,512	9,978.70	77,829
SB456bin23	Bioreactor19	131	3,356,086	2,666	25,619	174,375
SB456bin28	Bioreactor19	728	5,413,324	2,504	7,435.90	49,118
SB456bin35	Bioreactor19	221	1,727,919	2,622	7,818.60	36,518
SB456bin39	Bioreactor19	581	2,864,428	2,507	4,930.20	18,053
SB456bin42	Bioreactor19	384	2,022,548	2,510	5,267.10	29,960
SB456bin49	Bioreactor19	237	4,115,571	2,513	17,365.30	134,855
SB456bin50	Bioreactor19	533	3,581,737	2,501	6,720	29,378
SB460bin3	Bioreactor19	99	1,174,231	2,722	11,860.90	62,378
SB460bin7	Bioreactor19	83	6,511,571	2,825	78,452.70	441,255
SB508bin11	Bioreactor20	704	3,830,573	2,509	5,441.20	24,286
SB508bin15	Bioreactor20	44	3,356,849	10,053	76,292	284,020
SB508bin16	Bioreactor20	53	5,583,541	4,576	105,349.80	459,917
SB508bin20	Bioreactor20	365	5,593,220	2,524	15,323.90	120,391
SB508bin26	Bioreactor20	406	4,856,119	2,515	11,960.90	89,547
SB508bin36	Bioreactor20	313	1,908,331	2,503	6,096.90	22,658
SB508bin9	Bioreactor20	28	4,069,208	2,721	145,328.90	621,844
SB516bin19	Bioreactor20	886	5,000,757	2,501	5,644.20	30,408
SB516bin3	Bioreactor20	291	3,651,989	2,534	12,549.80	64,472
SB516bin30	Bioreactor20	463	3,265,432	2,503	7,052.80	30,733
SB516bin43	Bioreactor20	166	4,586,550	2,731	27,629.80	190,057
SB516bin5	Bioreactor20	1,497	7,292,330	2,500	4,871.30	25,091
SB516bin9	Bioreactor20	154	3,996,255	2,517	25,949.70	251,909
SB596bin10	Bioreactor20	585	7,360,122	2,511	12,581.40	97,695
SB596bin16	Bioreactor20	454	2,699,453	2,507	5,945.90	24,093
SB596bin29	Bioreactor20	268	3,477,774	2,520	12,976.80	78,993
SB596bin34	Bioreactor20	478	3,127,550	2,511	6,543	25,955
SB596bin39	Bioreactor20	314	3,368,312	2,531	10,727.10	87,766
SB604bin27	Bioreactor21	192	3,080,846	2,503	16,046.10	113,005
SB604bin5	Bioreactor21	267	3,017,095	2,503	11,300	45,270
SB931bin18	Bioreactor21	754	7,333,503	2,500	9,726.10	95,556
SB931bin3	Bioreactor21	222	4,797,199	2,530	21,609	147,562
SB935bin8	Bioreactor21	366	4,314,848	2,533	11,789.20	74,262
SG2bin19	Smugglers Gulch	229	1,605,672	2,511	7,011.70	47,578
SG2bin4	Smugglers Gulch	418	2,517,364	2,502	6,022.40	29,876
SG2bin6	Smugglers Gulch	184	2,599,350	2,587	14,126.90	72,703
SG3bin10	Smugglers Gulch	384	2,299,830	2,510	5,989.10	28,335

SRA1bin11	Salinas River Refuge	905	5,998,908	2,500	6,628.60	39,104
SRA1bin13	Salinas River Refuge	335	4,311,390	2,507	12,869.80	136,512
SRA1bin25	Salinas River Refuge	357	4,030,802	2,535	11,290.80	53,265
SRA1bin8	Salinas River Refuge	94	4,259,923	2,522	45,318.30	257,877
SRA3bin10	Salinas River Refuge	206	2,747,779	2,524	13,338.70	96,115
SRA3bin11	Salinas River Refuge	469	2,626,428	2,506	5,600.10	25,557

Genome Annotation

The resulting 60 high-quality dereplicated MAGs were then input into the DRAM tool in KBase for gene annotation using the KEGG Database (Kanehisa & Goto 2020). The resulting MAG annotations were filtered based on genes that were necessary to complete two KEGG metabolic pathways: denitrification and dissimilatory nitrate reduction to ammonia (DNRA). Some MAGs from the Salinas River Wildlife Refuge and the bioreactor had the presence of all genes needed to complete denitrification and DNRA (Figures 18 and 19).

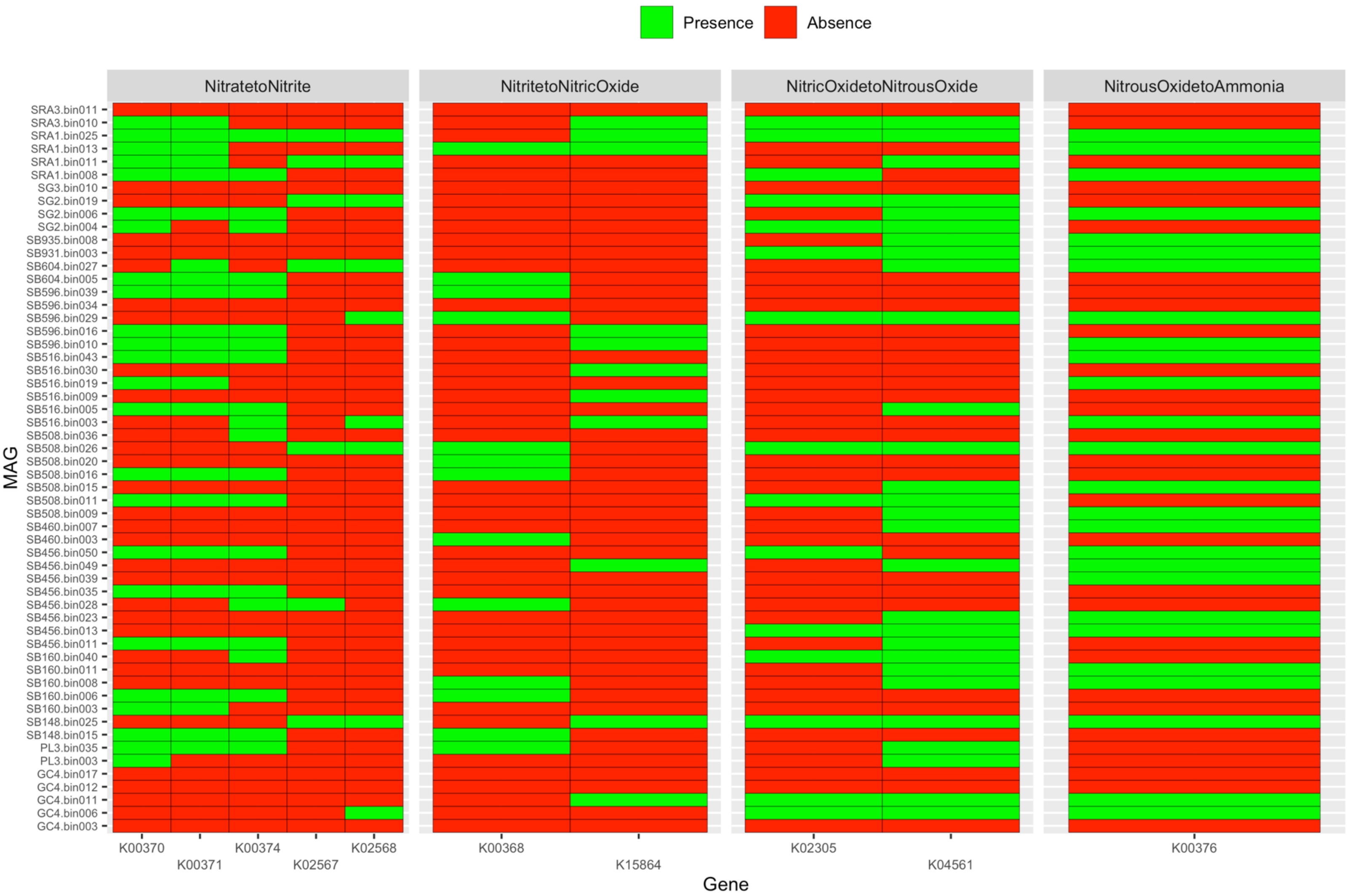

Figure 18. MAGs, from the 60 high-quality dereplicated MAGs, with at least 1 gene required for the denitrification pathway to be completed.

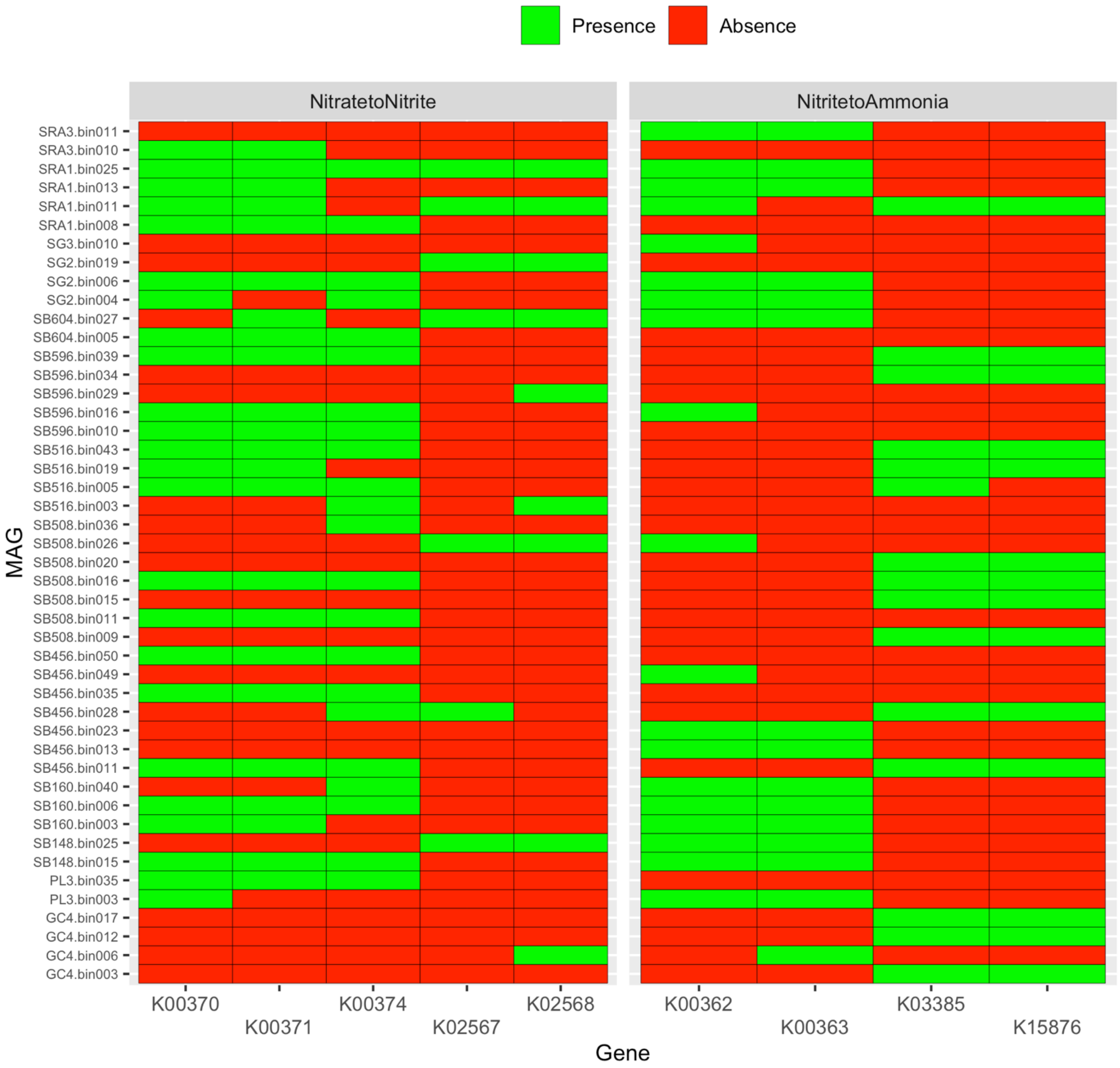

Figure 19. MAGs, from the 60 high-quality dereplicated MAGs, with at least 1 gene required for the DNRA pathway to be completed.

Taxonomic Assignments

The resulting 60 high-quality dereplicated MAGs were then input into the GTDB-Tk tool in KBase to assign taxonomies based on the comparison of all genes within a reference genome database made up to ~25,000 genomes. Overall, *Proteobacteria* was found in almost all location sites with the exception of Bioreactor21 (Figure 20). Within Smuggler's Gulch and Goat Canyon, the phyla *Campylobacterota* was unique when compared to the phyla in all other location sites (Figure 20). A similar relationship was seen with the phyla *Myxococcota* within the Bioreactor20 and the Salinas River National Wildlife Refuge (Figure 20).

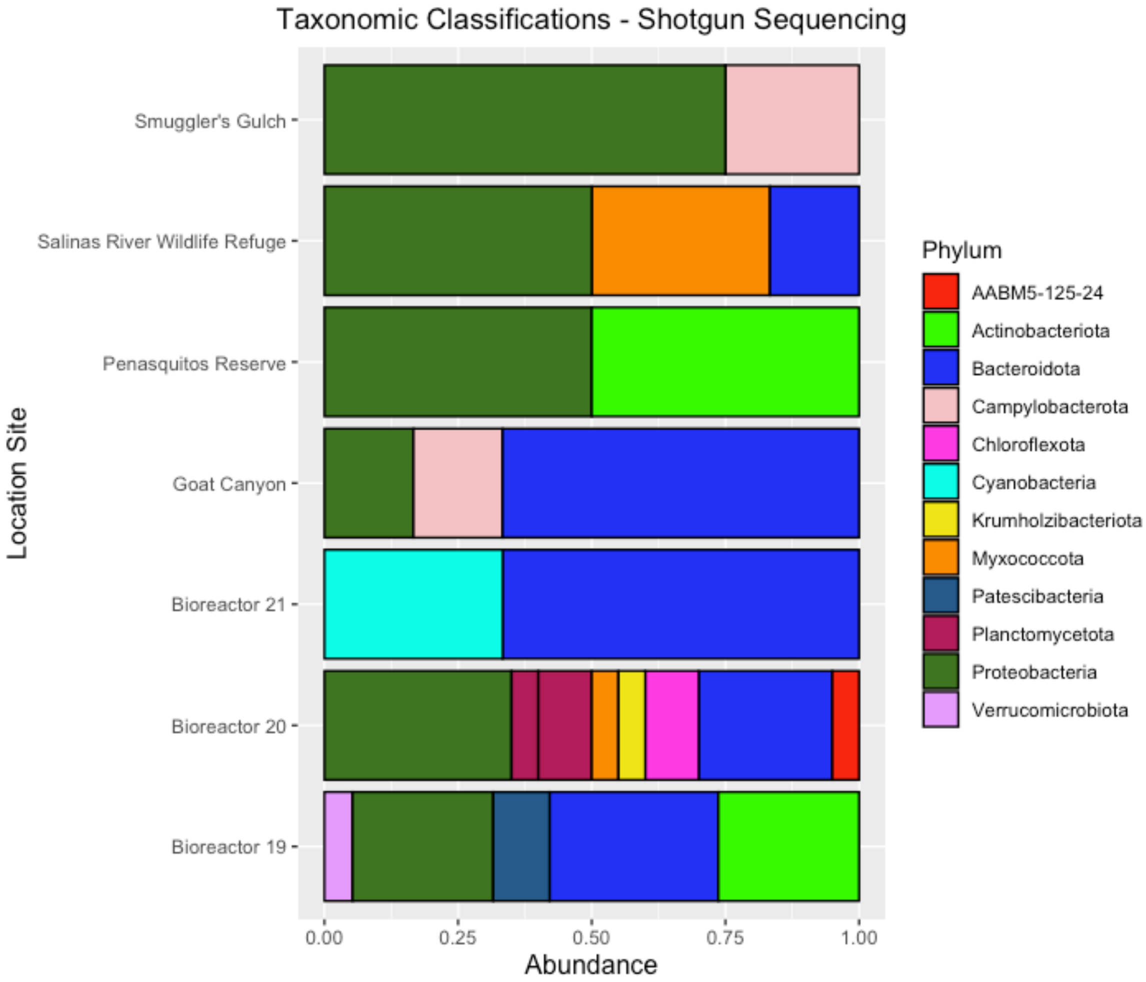

Figure 20. Phylum level relative abundance taxonomic assignments of the 60 high-quality dereplicated MAGs within each location site.

GO Term Assignments

The protein sequences found within the dereplicated high-quality bins were analyzed at Level 3 of the Biological Process GO category to assess metabolic function. The raw sequence counts within Level 3 can be found in Table A6 in the Appendix. Samples with total sequence counts above 2000 were assessed using relative abundance of sequence counts per GO category on a heatmap (Figure 21). The resulting dendrogram on the heatmap defined two main groups of samples: Group A and Group B (Table 24). The three categories that had the greatest representation within the samples were primary metabolic process, cellular metabolic process, and organic substance metabolic process (Figure 21). The three categories that were represented the least within the samples were regulation of metabolic process, signal transduction, and cell communication (Figure 21). The GO annotations from both Group A and Group B were then separated into two lists that were put through a two-tailed Fisher's Exact Test to find any significant differences between the two groups based on over or under representation of GO terms. If the proportion of genes annotated with a GO term in Group A was significantly higher than the proportion in the Group B, this GO term was be detected as overrepresented, and otherwise, it was declared underrepresented. Overall, there were 17 total GO terms that were underrepresented and 66 that were overrepresented (Table 25). After the Fisher's Exact Test was run on all the GO annotations, the GO terms found within the Level 3 Biological Process GO category were separated to further evaluate any trends in over or underrepresentation (Table 26). The complete list of over and underrepresented Level 3 GO category terms can be found in Tables A7 and A8 in the Appendix.

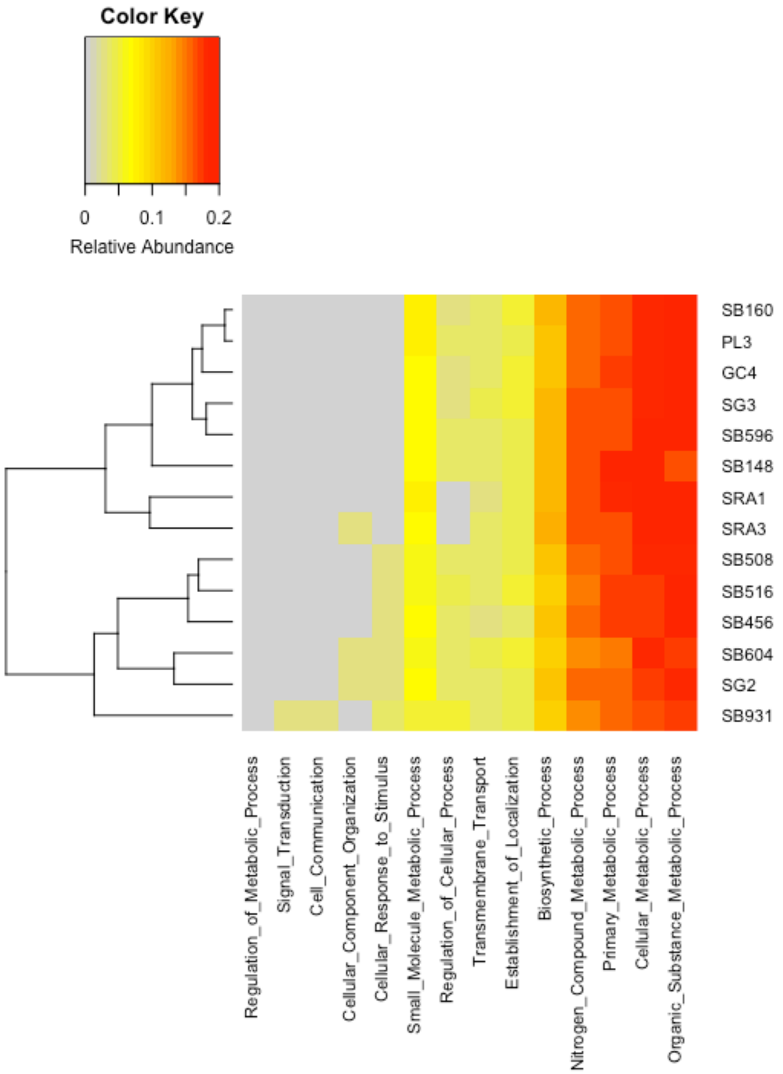

Figure 21. Relative abundance of Level 3 GO categories found within each sample with over 2000 sequences assigned to GO terms.

Table 24. List of samples within the two hierarchical groups defined by the dendrogram.

Dendrogram Groups	
Group A	Group B
SB160	SB508
PL3	SB516
GC4	SB456
SG3	SB604
SB596	SG2
SB148	SB931
SRA1	
SRA3	

Table 25. Results of the pair-wise Fisher's Exact Test in OmicsBox.

	Total Number of Sequences
Group A	107193
Group B	83363
Total Number of GO terms	2227
Go terms with "Over" Tag	17
Go terms with "Under" Tag	66

Table 26. Results of the pair-wise Fisher's Exact Test in OmicsBox for the Level 3 GO categories.

	Number of GO Terms	
Level 3 GO Category	Overrepresented	Underrepresented
Transmembrane Transport	3	0
Biosynthetic Process	4	9
Cellular Metabolic Process	0	1
Small Molecule Metabolic Process	0	0
Nitrogen Compound Metabolic Process	0	1
Organic Substance Metabolic Process	0	1
Primary Metabolic Process	0	2
Establishment of Localization	0	0
Regulation of Cellular Process	0	1
Cellular Response to Stimulus	0	1
Cell Communication	0	1
Signal Transduction	0	1
Cellular Component Organization	0	0
Regulation of Metabolic Process	0	1
Total	7	19

Discussion

Alpha Diversity, Environmental Conditions, and Taxonomic Assignments

The effects of coastal pollution on bacterial community structure have been well studied due to increasing anthropogenic activities taking place in nearby terrestrial locations. The presence of different types of pollution and environmental conditions has been shown to affect microbial diversity and significantly change the community composition (Mendez-Garcia et al. 2014; Xiong et al. 2012, Yuan et al. 2016; Dai et al. 2013; Hu et al. 2017). Within this study, different sampling location sites had a significant effect on alpha diversity (Observed and Shannon; Tables 11-14). Both diversity measures had the lowest and highest diversity values within Goat Canyon (Polluted Tijuana) and the CDPR Salinas Site (Polluted Salinas), respectively. Goat Canyon was also observed to have the lowest dissolved oxygen levels compared to all other location sites (Table A5).

Polluted Tijuana: Smuggler's Gulch and Goat Canyon

A 2018 study drafted by the U.S. Customs and Border Patrol showed several sampling locations within the Tijuana River Valley to have heavy metal concentrations above EPA regional screening levels (U.S. Customs and Border Patrol n.d.). Increasing levels of heavy metal contamination have been correlated to significant decreases in alpha diversity of bacteria found in soils (Qi et al. 2022). The lowest observed diversity values within two locations in the Tijuana River Valley, Smuggler's Gulch and Goat Canyon, could be explained by persisting high levels of heavy metals within the soil. In addition to heavy metals, volatile organic compounds (VOCs) have also been observed within the Tijuana River Valley (U.S. Customs and Border Patrol n.d.). According to

Abis et al. 2020, VOC emissions in soils are positively correlated to two bacterial phyla: *Proteobacteria and Bacteriodetes*. Samples collected within this study from both Goat Canyon and Smuggler's Gulch had a presence of the phylum *Proteobacteria*, which could be an indication of persisting high levels of VOCs in Tijuana River Valley soils, but data were not located that could verify this hypothesis.

Smuggler's Gulch and Goat Canyon were also observed to have the presence of a unique phylum not found in any of the other sampling locations: *Campylobacterota*. *Campylobacterota* can survive in a diverse array of habitats ranging from marine ecosystems to internal animal organs (Eppinger et al. 2004; Waite et al. 2017; Parks et al. 2018). Within the *Campylobacterota* phylum, the family *Arcobacteraceae* was observed within Smuggler's Gulch and Goat Canyon. The phyla *Campylobacterota,* and the family *Arcobacteraceae,* have both been associated with human and animal illnesses (Venâncio et al. 2022; van der Stel and Wösten 2019). In Venâncio et al. 2022, the geographic distribution of *Arcobacteraceae* was studied to better understand any health risks due to the presence of the taxonomic family. Bacteria within the family *Arcobacteraceae* had the highest prevalence within raw sewage and wastewater when compared to seawater, surface water, groundwater, and water processed in food processing plants (Venâncio et al. 2022). The presence of this family within the two sampling locations in the Tijuana River Valley could suggest a presence of raw sewage or wastewater conditions similar to those studied in Venâncio et al. 2022. Further environmental monitoring and chemical analyses should be performed to confirm this potential relationship.

Polluted Salinas: CDPR Salinas Site

The CDPR monitoring site along the Salinas River at Davis Road had the highest alpha diversity levels within all sampling locations (Table 10). According to a water quality monitoring report drafted in 2020, the CDPR monitoring site was observed to have acute and chronic concentrations of three different common pesticides (CCRWQCB 2020). According to Onwona-Kwakye et al. 2020, a presence of pesticides within soil corresponded to lower levels of bacterial diversity and composition, with specific genera increasing or decreasing based on pesticide exposure. Top genera found within the CDPR monitoring site were not related to those that increased or decreased based on pesticide exposure within Onwona-Kwakye et al. 2020.

Beta Diversity and Ordination Methods

All sampling location sites, as well as nitrate and phosphate concentrations, pH, dissolved oxygen, temperature, and turbidity, were observed to significantly affect Bray-Curtis beta diversity values. This is consistent with current literature which shows significant differences in beta diversity between different sampling sites based on physical and chemical variability within the environment (Hengy, et al. 2017).

Within this study, the NMDS was calculated with 3 axes and had a resulting stress score below 0.1 (0.08). According to foundational literature (Clarke 1993), stress values less than 0.1 correspond to appropriate ordination with “no real risk of drawing false inferences.” Clarke 1993 emphasized the importance of using Shepard and scree plots to further validate the ordination fit in addition to the stress value. These plots were analyzed within this study showing appropriate fit based on R^2 values. These analyses result in a reliable NMDS ordination of dissimilarity values for this study. The NMDS ordination showed clustering based on sampling location site. This clustering is expected since PERMANOVA values showed a significant difference in beta diversity due to sampling location site. The NMDS ordination also showed three larger clusters based on comparison categories: bioreactor, polluted Tijuana, and polluted Salinas which could be explained by the PERMANOVA showing significant differences in beta diversity based on comparison categories.

Limitations of Ordination Methods

Because of the relatively low rarefaction sampling depth applied to all samples when compared to the high number of unique ASVs, there were many zeros seen within the working data set of ASVs. When PCA was used within this study, a “horseshoe” visualization effect was observed within the plot, as described by other studies when running this same statistical analysis on datasets with similar high zero abundances (Legendre & Legendre 2012; ter Braak & Šmilauer 2014). Although there was high variance explained by the first two PCs, PCA would not be an appropriate method of displaying beta diversity within this study unless certain transformations are applied to the original dataset (Paliy & Shankar 2016). These transformations were not explored within this study.

PCoA is different from NMDS in that it attempts to maximize a linear relationship of ASV abundances along a gradient, which can also result in a “horseshoe” visualization similar to that of PCA with high zero values within a working dataset (Podani & Miklós, 2002). A “horseshoe” visualization can also be observed within the resulting PCoA ordination of this study, suggesting an incorrect ordination of dissimilarity values.

Presence of Metabolic Pathways of Interest

Denitrification

Within all high-quality dereplicated MAGs, no singular MAG had all genes required to complete denitrification. Because microbes live within communities in their environment, different species often rely on each other to complete metabolic pathways and produce energy (Kouzuma et al. 2015). Through a mutualistic process known as syntrophy, microbes excrete metabolites that promote growth within other microbes in the same community (Kouzuma et al. 2015). This process has been seen within natural and artificial ecosystems, giving unique microbes the ability to survive within their environment (Kouzuma et al. 2015). With the understanding of syntrophy, all genes within a sampling location site were analyzed together. When combining all genes from all MAGs of a sampling location site, the two sites that had all necessary genes to complete denitrification were the Salinas River Wildlife Refuge and the bioreactor. Denitrification rates have been significantly affected by nitrate concentrations and temperature of water within the environment (Tomasek et al. 2017; Allin et al. 2017).

Within this study, the bioreactor had the highest levels of nitrate concentrations and temperature when compared to all other sampling locations (Table A5). Because these conditions favor denitrification rates, the bioreactor could be promoting increased levels of denitrification due to its closed system that allows for environmental conditions to be minimally influenced by outside factors. In order to validate the levels of denitrification within this study, an environment-specific method of measuring denitrification could be used to estimate and compare denitrification rates among all sampling location sites (Groffman et al. 2006). Because most other sampling locations within this study have evidence of incomplete denitrification to ammonia based on the absence of denitrifying genes, nitric oxide and nitrous oxide levels should also be measured to quantify any affect they may have on greenhouse gas emissions of the location (Tomasek et al. 2017; Ravishankara et al. 2009).

The second sampling location site that had a presence of all necessary genes to complete denitrification is the Salinas River National Wildlife Refuge. The Salinas River National Wildlife Refuge had the second lowest nitrate values out of all sampling location sites which does not follow the relationship between nitrates and denitrification outlined in Tomasek et al. 2017. According to current literature, other factors that significantly affect denitrification of soils include organic carbon content, soil moisture, oxygen concentrations, and water movement within an environment (Perryman et al., 2011; Inwood et al., 2007; Pinay et al., 2007; O'Connor & Hondzo, 2008). These factors could be measured to better assess the status of denitrification within this sampling location.

The completion of denitrification within the Salinas River National Wildlife Refuge could also be influenced by hydrologic connectivity of the location. Hydrologic connectivity is the transfer of energy within an ecosystem through the hydrologic cycle (Pringle 2003). Increased biogeochemical cycling has been linked to environments characterized as having greater hydrologic connectivity which has been linked to increased rates of denitrification (Wantzen & Junk 2006; McClain et al. 2003). In order to better understand this relationship, a future study could quantify hydrologic connectivity within all sites to further investigate any factors affecting denitrification.

Dissimilatory Nitrate Reduction to Ammonia (DNRA)

Within all high-quality dereplicated MAGs, no singular MAG had all genes required to complete DNRA. When combining all genes from all MAGs of a sampling location site, the two sites that had all necessary genes to complete DNRA were the Salinas River National Wildlife Refuge and the bioreactor. DNRA in sediments has been shown to be affected by various environmental factors including nitrates, iron, sulfide, organic carbon, temperature, pH, and precipitation (Dong et al. 2011; Laverman et a. 2007; Cheng et al. 2022). Current literature shows that precipitation is the main driver of DNRA in soil (Cheng et al. 2022). Precipitation was not measured within sampling locations but should be considered within future studies to better analyze DNRA completion within a location. In a study by van den Berg et al. 2016, microbes responsible for DNRA were observed to increase when limited amounts of nitrate were also observed within the environment. This trend could explain the completion of DNRA within the Salinas River National Wildlife Refuge since the nitrate concentrations within the location site were some of the lowest compared to all other sampling locations. Similar to the relationship between temperature and denitrification, DNRA has also been observed to increase with rising temperatures within different environments (Lai et al. 2021). Similar to the relationship between temperature and denitrification within the bioreactor, the completion of DNRA could be due to higher temperatures within the closed system when compared to all other sampling locations.

Taxonomic Classifications Related to Nitrate Reduction – 16S rRNA and shotgun metaG

Bacterial phyla that are most well-known for reducing nitrate within the environment include *Proteobacteria, Bacteroidota, Bacteroidetes, Firmicutes,* and *Actinobacteria* (Jones et al. 2008; Nelson et al. 2016). Analyzing 16S rRNA taxonomic classifications, *Actinobacteria and Proteobacteria* were two of the top phyla observed within all sampling location sites. Proteobacteria had a higher abundance within the bioreactor system, which could explain the presence of both metabolic nitrate reduction pathways found within the system. The bacterial phylum most abundant within the Salinas River National Wildlife Refuge was observed to be *Actinobacteria,* but this trend was also observed within other location sites. Analyzing whole metagenome shotgun taxonomic classifications, *Proteobacteria, Actinobacteria*, and *Bacteroidota* were present within almost all sampling location sites, with higher frequencies observed within the bioreactor and the Salinas River National Wildlife Refuge. A unique phylum was found within the bioreactor and the Salinas River National Wildlife Refuge and was classified as *Myxococcota*. Within current literature, cultured *Myxococcota* have been categorized as aerobic soil microbes that have unique predation strategies (Murphy et al. 2021). Uncultured *Myxococcota* found within anaerobic environments have been observed to have smaller genomes with less genes than their cultured counterparts (Murphy et al. 2021). These unculture *Myxococcota* have also been observed to exclusively rely on fermentation and nitrate and sulfate reduction to produce energy due to evolution of predation in changing soils (Murphy et al. 2021). The presence of *Myxococcota* within the bioreactor and the Salinas River National Wildlife Refuge could explain the completion of both nitrate reducing pathways and evidence of evolution of predation within the uncultured phylum.

GO Term Annotations

The dendrogram produced within the heatmap showing relative abundance of sequences within the Level 3 GO categories produced two main hierarchical groups. Group A consisted of sequences from samples collected from Bioreactor 2019, Penasquitos Reserve, Goat Canyon, Smuggler's Gulch, Bioreactor 2020, and the Salinas River National Wildlife Refuge. Group B consisted of sequences from samples collected from Bioreactor 2020, Bioreactor 2019, Smuggler's Gulch, and Bioreactor 2021. Goat Canyon, Penasquitos Reserve, and the Salinas River National Wildlife Refuge were the locations not represented by the samples in Group B but were present in Group A. Group B included sequences from samples collected from Bioreactor 2019 that were not represented in Group A. Level 3 GO category terms within the different sample locations were analyzed to understand the functional potential of specific GO terms.

Group A had overrepresentation of only two Biological Process Level 3 Categories: Transmembrane Transport and Biosynthetic Process (Table 26). The Transmembrane Transport Level 3 GO category was the only category that was overrepresented while also not being underrepresented by samples within Group A. The GO term most overrepresented within the Transmembrane Transport category was Amino Acid Transmembrane Transport (p=6.81E-04, GO:0003333, Table A7). Amino Acid Transmembrane Transport involves the movement of an amino acid across a cell membrane (Carbon et al. 2009). Within microbial systems, amino acids are responsible for maintaining cell structure, as well as the internal environment of the organism to allow it to survive (Zeden et al 2021). In a recent study, a specific bacterium was observed to change its amino acid concentrations as a response to different environmental factors, specifically pH and oxygen content (Zeden et al. 2021). Overrepresentation of Amino Acid Transmembrane Transport could be an indication of significant differences in environmental conditions between location sites represented by the samples in Group A and Group B. Included only within Group A, Salinas River National Wildlife Refuge, Penasquitos Reserve, and Goat Canyon location sites had the highest pH observations when compared to the other location sites represented by Group B. Increased pH within these location sites could be driving the bacteria to extract more amino acids from the environment, as seen similarly within *S. aureus* in Zhu et al. 2007. Coupled with the adaptive strategy to acquire specific amino acids from the environment, bacteria also redirect energy necessary for growth from Amino Acid Biosynthesis (Zu et al. 2007), another GO term that was overrepresented within Group A (GO:0008652, p=3.74E-05, Table A7). Similar to *S. aureus*, the phylum *Campylobacterota* found only within Goat Canyon and Smuggler's Gulch is associated with human and animal illnesses (Venâncio et al. 2022; van der Stel and Wösten 2019). The MAGs assigned as *Campylobacterota* in this study, from both Goat Canyon and Smuggler's Gulch, could be investigated further to understand the specific amino acids extracted from the environment. This could give more information explaining the overrepresentation of the Transmembrane Transport Level 3 GO category.

Group A had underrepresentation of 8 Biological Process Level 3 Categories: Cell Communication, Regulation of Cellular Process, Cellular Response to Stimulus, Primary Metabolic Process, Biosynthetic Process, Organic Substance Metabolic Process, and

Cellular Metabolic Process (Table 26). The most underrepresented GO term in Group A was cell communication (GO:0007154, p=4.12E-41, Table A8). Cell-to-cell communication has also been referred to as quorum sensing and has been observed to affect the expression profile of particular genes within a microbial community (Hirakawa and Tomita 2013). Quorum sensing occurs with the help of "autoinducers" or chemical compounds that drive and suppress certain behaviors in bacteria (Hirakawa and Tomita 2013). Quorum sensing was first studied in the context of bioluminescence (Nealson et al. 1970) but is now extensively being studied within pathogenic bacteria and has led to the theory that bacteria might be regulating their viral genes (Hirakawa and Tomita 2013). Although Group B had no location sites represented that completed the two pathways of interest within this study, the introduction of genes responsible for completing the pathways could be regulated and expressed with the help of quorum sensing that was overrepresented in the group. Quorum sensing is also related to Regulation of Cellular Process (GO:0050794), which was the second-most underrepresented GO term in Group A (overrepresentation in Group B, p= 3.82E-28, Table A8), and involves any modification of frequency, rate, or extent of cellular processes that occur at the cellular level and are not limited to single-cell activity (Carbon et al. 2009). Different environmental features have been observed to affect cell-to-cell communication (Mukherjee and Bassler et al. 2019) and should continue to be studied to understand how bacteria will respond to changing environments related to pollution.

Conclusions and Future Directions

This study reveals significant differences in bacterial diversity and community composition within sampling locations in two historically polluted coastal environments in California. Within sampling locations of this study beta diversity was affected by nitrate, phosphate, dissolved oxygen, pH, temperature, and turbidity. The presence of two complete nitrate-reducing metabolic pathways were found within two sampling locations: the bioreactor and the Salinas River National Wildlife Refuge. Nitrate and temperature observations, in addition to the presence of specific bacterial phyla, within these sampling locations could explain the completion of the nitrate reducing metabolic pathways when compared to all other sampling locations.

Other immediate analyses that could be completed to further understand the trends seen within this study include metagenomic read mapping and gene expression profiling. Read mapping allows for the relative abundance of each individual MAG to be calculated within every metagenomic sample (Desai et al. 2013). This can give insight into which specific MAGs could be contributing to the nitrate-reducing metabolic potential of the different sampling locations based on gene relative abundances (Sharpton 2014). Gene expression is analyzed within microbial studies to understand the synthesis of the product of a functional gene (Bervoets 2019). Incorporating this analysis within a future study could provide information on the use of nitrate reducing genes within microbes found in the sampling locations and their roles within the specific environment.

References

Abis, L., Loubet, B., Ciuraru, R., Lafouge, F., Houot, S., Nowak, V., Tripied, J., Dequiedt, S., Maron, P. A., & Sadet-Bourgeteau, S. 2020. Reduced microbial diversity induces larger volatile organic compound emissions from soils. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-63091-8

Allin, A., Schernewski, G., Friedland, R., Neumann, T., & Radtke, H. 2017. Climate change effects on denitrification and associated avoidance costs in three Baltic River basin - coastal sea systems. Journal of Coastal Conservation, 21(4), 561–569. https://doi.org/10.1007/s11852-017-0530-8

Al-Shahrour F., Díaz-Uriarte R. and Dopazo J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics (Oxford, England), 20(4), 578-80.

Andrews, S. 2010. FASTQC. A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

van den Berg, E. M., Boleij, M., Kuenen, J. G., Kleerebezem, R., & van Loosdrecht, M. C. 2016. DNRA and denitrification coexist over a broad range of acetate/n-NO3– ratios, in a chemostat enrichment culture. Frontiers in Microbiology, 7. https://doi.org/10.3389/fmicb.2016.01842

Bervoets, I., & Charlier, D. 2019. Diversity, versatility and complexity of bacterial gene regulation mechanisms: Opportunities and drawbacks for applications in Synthetic Biology. FEMS Microbiology Reviews, 43(3), 304–339. https://doi.org/10.1093/femsre/fuz001

BioBam Bioinformatics. 2019. OmicsBox - Bioinformatics made easy (Version 3.0.29). Retrieved at: www.biobam.com/omicsbox.

Bolyen, E., Rideout, J.R., Dillon, M.R. et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol, 37, 852–85. https://doi.org/10.1038/s41587-019-0209-9

Bu, C., Wang, Y., Ge, C., Ahmad, H. A., Gao, B., & Ni, S.-Q. 2017. Dissimilatory nitrate reduction to ammonium in the Yellow River estuary: Rates, abundance, and community diversity. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-06404-8

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics, 25(2), 288-289.

[CCRWQCB] Central Coast Regional Water Quality Control Board. 2020. Draft: Total Maximum Daily Loads to Address Organophosphate Pesticides and Aquatic

Toxicity Impairments within the Lower Salinas River Watershed. Water Quality Analysis Report. Retrieved from: https://www.waterboards.ca.gov/centralcoast/water_issues/programs/tmdl/docs/salinas/oppesticides/docs/draft_wqda_rpt.pdf

ter Braak, C. J., & Šmilauer, P. 2014. Topics in constrained and unconstrained ordination. Plant Ecology, 216(5), 683–696. https://doi.org/10.1007/s11258-014-0356-5

Chandran, H., Meena, M., & Sharma, K. 2020. Microbial Biodiversity and bioremediation assessment through OMICS approaches. Frontiers in Environmental Chemistry, 1. https://doi.org/10.3389/fenvc.2020.570326

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.J. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol Lett 8(2):148–159

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. 2019. GTDB-TK: A toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz848

Cheng, Y., Elrys, A. S., Merwad, A.-R. M., Zhang, H., Chen, Z., Zhang, J., Cai, Z., & Müller, C. 2022. Global Patterns and drivers of soil dissimilatory nitrate reduction to ammonium. Environmental Science & Technology, 56(6), 3791–3800. https://doi.org/10.1021/acs.est.1c07997

Chouinard, O., Jorgensen, B., Tett, P., Vanderlinden, J.-P., Vasseur, L., Baztan, J., & Wright, W. W. 2015. Coastal zones: Solutions for the 21st Century. Elsevier.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Austral Ecology, 18(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Clarridge, J. E. 2004. IMPACT OF 16S rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews, 17*(4), 840–862. https://doi.org/10.1128/cmr.17.4.840-862.2004

Dai, J., Tang, X., Gao, G., Chen, D., Shao, K., Cai, X., & Zhang, L. 2013. Effects of salinity and nutrients on sedimentary bacterial communities in Oligosaline Lake Bosten, Northwestern China. Aquatic Microbial Ecology, 69(2), 123–134. https://doi.org/10.3354/ame01627

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., & Jere, A. 2013. Identification of optimum sequencing depth especially for de Novo genome assembly of small genomes using Next Generation Sequencing Data. PLoS ONE, 8(4). https://doi.org/10.1371/journal.pone.0060204

Dong, L. F., Sobey, M. N., Smith, C. J., Rusmana, I., Phillips, W., Stott, A., Osborn, A. M., & Nedwell, D. B. 2010. Dissimilatory reduction of nitrate to ammonium, not denitrification or anammox, dominates benthic nitrate reduction in tropical estuaries. Limnology and Oceanography, 56(1), 279–291. https://doi.org/10.4319/lo.2011.56.1.0279

Dow, E. G., Wood-Charlson, E. M., Biller, S. J., Paustian, T., Schirmer, A., Sheik, C. S., Whitham, J. M., Krebs, R., Goller, C. C., Allen, B., Crockett, Z., & Arkin, A. P. 2021. Bioinformatic teaching resources – for educators, by educators – using KBase, a free, user-friendly, open source platform. Frontiers in Education, 6. https://doi.org/10.3389/feduc.2021.711535

Dowd, B., Press, D., & Huertos, M. 2008. Agricultural Nonpoint Source Water Pollution Policy: The case of California's Central Coast. Agriculture, Ecosystems & Environment, 128(3), 151–161. https://doi.org/10.1016/j.agee.2008.05.014

Eppinger, M., Baar, C., Raddatz, G., Huson, D. H., & Schuster, S. C. 2004. Comparative analysis of four Campylobacterales. Nature Reviews Microbiology, 2(11), 872–885. https://doi.org/10.1038/nrmicro1024

Fadeev, E., Cardozo-Mino, M. G., Rapp, J. Z., Bienhold, C., Salter, I., Salman-Carvalho, V., Molari, M., Tegetmeyer, H. E., Buttigieg, P. L., & Boetius, A. 2021. Comparison of two 16S rrna primers (v3–v4 and V4–v5) for studies of Arctic Microbial Communities. Frontiers in Microbiology, 12. https://doi.org/10.3389/fmicb.2021.637526

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., & Ravel, J. 2014. An improved dual-indexing approach for multiplexed 16S rrna gene sequencing on the Illumina MiSeq Platform. Microbiome, 2(1). https://doi.org/10.1186/2049-2618-2-6

Götz, S., Garcia-Gomez, J.M., Terol J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-35.

Groffman, P. M., Altabet, M. A., Böhlke, J. K., Butterbach-Bahl, K., David, M. B., Firestone, M. K., Giblin, A. E., Kana, T. M., Nielsen, L. P., & Voytek, M. A. 2006. Methods for measuring denitrification: Diverse approaches to a difficult problem. Ecological Applications, 16(6), 2091–2122. https://doi.org/10.1890/1051-0761(2006)016[2091:mfmdda]2.0.co;2

Hengy, M. H., Horton, D. J., Uzarski, D. G., & Learman, D. R. 2017. Microbial community diversity patterns are related to physical and chemical differences among temperate lakes near Beaver Island, MI. PeerJ, 5. https://doi.org/10.7717/peerj.3937

Hicks, N., Liu, X., Gregory, R., Kenny, J., Lucaci, A., Lenzi, L., Paterson, D. M., & Duncan, K. R. 2018. Temperature driven changes in benthic bacterial diversity influences biogeochemical cycling in coastal sediments. Frontiers in Microbiology, 9. https://doi.org/10.3389/fmicb.2018.01730

Hirakawa, H., & Tomita, H. 2013. Interference of bacterial cell-to-cell communication: A new concept of antimicrobial chemotherapy breaks antibiotic. Frontiers in Microbiology, *4*. https://doi.org/10.3389/fmicb.2013.00114

Hu, A., Ju, F., Hou, L., Li, J., Yang, X., Wang, H., Mulla, S. I., Sun, Q., Bürgmann, H., & Yu, C.-P. 2017. Strong impact of anthropogenic contamination on the co-occurrence patterns of a riverine microbial community. Environmental Microbiology, 19(12), 4993–5009.

Inwood, S. E., Tank, J. L., and Bernot, M. J. 2007. Factors controlling sediment denitrification in midwestern streams of varying land use. Microb. Ecol. 53, 247–258. doi: 10.1007/s00248-006-9104-2https://doi.org/10.1111/1462-2920.13942

Janda, J. M., & Abbott, S. L. 2007. 16S rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and Pitfalls. Journal of Clinical Microbiology, 45(9), 2761–2764. https://doi.org/10.1128/jcm.01228-07

Jones, C. M., Stres, B., Rosenquist, M., & Hallin, S. 2008. Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification. Molecular Biology and Evolution, 25(9), 1955–1966. https://doi.org/10.1093/molbev/msn146

Kamp, A., Høgslund, S., Risgaard-Petersen, N., & Stief, P. 2015. Nitrate storage and dissimilatory nitrate reduction by eukaryotic microbes. Frontiers in Microbiology, 6. https://doi.org/10.3389/fmicb.2015.01492

Kanehisa, M. & Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30. https://academic.oup.com/nar/article/28/1/27/2384332

Kang, D., Li, F., Kirton, E. S., Thomas, A., Egan, R. S., An, H., & Wang, Z. 2019. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from Metagenome Assemblies. https://doi.org/10.7287/peerj.preprints.27522v1

Kouzuma, A., Kato, S., & Watanabe, K. 2015. Microbial interspecies interactions: Recent findings in syntrophic consortia. Frontiers in Microbiology, 6. https://doi.org/10.3389/fmicb.2015.00477

Kuenen, J. G. 2008. Anammox bacteria: from discovery to application. Nat. Rev. Microbiol. 6, 320–326. doi: 10.1038/nrmicro1857

Lai, T. V., Ryder, M. H., Rathjen, J. R., Bolan, N. S., Croxford, A. E., & Denton, M. D. 2021. Dissimilatory nitrate reduction to ammonium increased with rising temperature. Biology and Fertility of Soils, 57(3), 363–372. https://doi.org/10.1007/s00374-020-01529-x

Laverman, A. M., Canavan, R. W., Slomp, C. P., & Cappellen, P. V. 2007. Potential nitrate removal in a coastal freshwater sediment (Haringvliet Lake, the Netherlands) and response to salinization. Water Research, 41(14), 3061–3068. https://doi.org/10.1016/j.watres.2007.04.002

Legendre, P., & Legendre, L. 2012. Numerical ecology. Elsevier Science.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10-12. https://doi.org/10.14806/ej.17.1.200

Maziarz, M., Pfeiffer, R. M., Wan, Y., & Gail, M. H. 2018. Using standard microbiome reference groups to simplify beta-diversity analyses and facilitate independent validation. Bioinformatics, *34*(19), 3249–3257. https://doi.org/10.1093/bioinformatics/bty297

McClain, M. E., Boyer, E. W., Dent, C. L., Gergel, S. E., Grimm, N. B., Groffman, P. M., et al. 2003. Biogeochemical hot spots and hot moments at the interface of terrestrial and Aquatic Ecosystems. Ecosystems 6, 301–312. doi: 10.1007/s10021-003-0161-9

McMurdie, P.J., & Holmes, S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217

Méndez-García, C., Mesa, V., Sprenger, R. R., Richter, M., Diez, M. S., Solano, J., Bargiela, R., Golyshina, O. V., Manteca, Á., Ramos, J. L., Gallego, J. R., Llorente, I., Martins dos Santos, V. A. P., Jensen, O. N., Peláez, A. I., Sánchez, J., & Ferrer, M. 2014. Microbial stratification in low ph oxic and suboxic macroscopic growths along an acid mine drainage. The ISME Journal, 8(6), 1259–1274. https://doi.org/10.1038/ismej.2013.242

Miller, G. A., & Chapman, J. P. 2001. Misunderstanding analysis of covariance. Journal of Abnormal Psychology, *110*(1), 40–48. https://doi.org/10.1037/0021-843x.110.1.40

Mortensen, Z., Kato, J., Silveus, J., & Valdez, A., & Hall, S., Nimmers, K., Haffa, A. 2019. Isolation of Microbial Populations with the Ability To Use Pesticides as a Sole Carbon Source in Multichannel Woodchip Bioreactors under a Controlled Environment. 10.1021/bk-2019-1308.ch024.

Mueller, J. G., Cerniglia, C. E., & Pritchard, P. H. 1996. Bioremediation of environments contaminated by polycyclic aromatic hydrocarbons. Bioremediation: Principles and Applications, eds Crawford R. L., Crawford L. D. (Cambridge: Cambridge University Press;), 125–194.

Mukherjee, S., & Bassler, B. L. 2019. Bacterial quorum sensing in complex and Dynamically Changing Environments. Nature Reviews Microbiology, *17*(6), 371–382. https://doi.org/10.1038/s41579-019-0186-5

Murphy, C. L., Yang, R., Decker, T., Cavalliere, C., Andreev, V., Bircher, N., Cornell, J., Dohmen, R., Pratt, C. J., Grinnell, A., Higgs, J., Jett, C., Gillett, E., Khadka, R., Mares, S., Meili, C., Liu, J., Mukhtar, H., Elshahed, M. S., & Youssef, N. H. 2021. Genomes of novel Myxococcota reveal severely curtailed machineries for predation and cellular differentiation. https://doi.org/10.1101/2021.07.06.451402

Nealson, K. H., Platt, T., Hastings, J. W. 1970. Cellular control of the synthesis and activity of the bacterial luminescent system. J. Bacteriol, 104, 313–322.

Nelson, M. B., Martiny, A. C., & Martiny, J. B. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. Proceedings of the National Academy of Sciences, 113(29), 8033–8040. https://doi.org/10.1073/pnas.1601070113

Nguyen, N. H., Smith, D., Peay, K., & Kennedy, P. 2014. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytologist, 205(4), 1389–1393. https://doi.org/10.1111/nph.12923

Nunes, M., & Leston, S. 2020. Coastal pollution: An overview. Encyclopedia of the UN Sustainable Development Goals, 1–11. https://doi.org/10.1007/978-3-319-71064-8_9-1

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. 2017. Metaspades: A new versatile metagenomic assembler. Genome Research, 27(5), 824–834. https://doi.org/10.1101/gr.213959.116

O'Connor, B. L., and Hondzo, M. 2008. Enhancement and inhibition of denitrification by fluid-flow and dissolved oxygen flux to stream sediments. Environ. Sci. Technol. 42, 119–125. doi: 10.1021/es071173s

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. 2017. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. The ISME Journal, 11(12), 2864–2868. https://doi.org/10.1038/ismej.2017.126

Onwona-Kwakye, M., Plants-Paris, K., Keita, K., Lee, J., Brink, P. J., Hogarh, J. N., & Darkoh, C. 2020. Pesticides decrease bacterial diversity and abundance of irrigated rice fields. Microorganisms, 8(3), 318. https://doi.org/10.3390/microorganisms8030318

Paliy, O., & Shankar, V. 2016. Application of multivariate statistical techniques in Microbial Ecology. Molecular Ecology, 25(5), 1032–1057. https://doi.org/10.1111/mec.13536

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology, 36(10), 996–1004. https://doi.org/10.1038/nbt.4229

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. 2014. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25: 1043-1055.

Pérez-Cobas, A. E., Gomez-Valero, L., & Buchrieser, C. 2020. Metagenomic approaches in Microbial Ecology: An update on whole-genome and marker gene sequencing analyses. Microbial Genomics, 6(8). https://doi.org/10.1099/mgen.0.000409

Perryman, S. E., Rees, G. N., Walsh, C. J., and Grace, M. R. 2011. Urban stormwater runoff drives denitrifying community composition through changes in sediment texture and carbon content. Microb. Ecol. 61, 932–940. doi: 10.1007/s00248-011-9833-8

Pinay, G., Gumiero, B., Tabacchi, E., Gimenez, O., Tabacchi-Planty, A. M., Hefting, M. M., et al. 2007. Patterns of denitrification rates in European alluvial soils under various hydrological regimes. Freshw. Biol. 52, 252–266. doi: 10.1111/j.1365-2427.2006.01680.x

Podani, J., & Miklós, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology, 83(12), 3331–3343. https://doi.org/10.1890/0012-9658(2002)083[3331:rcathe]2.0.co;2

Pringle, C. 2003. What is hydrologic connectivity and why is it ecologically important? Hydrological Processes, 17(13), 2685–2689. https://doi.org/10.1002/hyp.5145

Qi, Q., Hu, C., Lin, J., Wang, X., Tang, C., Dai, Z., & Xu, J. 2022. Contamination with multiple heavy metals decreases microbial diversity and favors generalists as the keystones in microbial occurrence networks. Environmental Pollution, 306, 119406. https://doi.org/10.1016/j.envpol.2022.119406

Raffa, C.M., & Chiampo, F. 2021. Bioremediation of Agricultural Soils Polluted with Pesticides: A Review. Bioengineering, 8(7), 92. https://doi.org/10.3390/bioengineering8070092

Ravishankara, A. R., Daniel, J. S., & Portmann, R. W. 2009. Nitrous oxide (N 2 O): The dominant ozone-depleting substance emitted in the 21st Century. Science, 326(5949), 123–125. https://doi.org/10.1126/science.1176985

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. 2020. Rescript: Reproducible sequence taxonomy reference database management for the Masses. https://doi.org/10.1101/2020.10.05.326504

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitua, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. 2020. Dram for distilling microbial metabolism to automate the curation of microbiome function. https://doi.org/10.1101/2020.06.29.177501

Shapleigh, J. P. 2006. "The denitrifying prokaryotes," in The Prokaryotes, eds M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schlieger, and E. Stackebrant (New York, NY: Springer), 769–792.

Sharma, S. 2012. Bioremediation: Features, Strategies and applications. 2, 12.

Sharpton, T. J. 2014. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science, 5. https://doi.org/10.3389/fpls.2014.00209

Stein, L. J., Gunier, R. B., Harley, K., Kogut, K., Bradman, A., & Eskenazi, B. 2016. Early childhood adversity potentiates the adverse association between prenatal organophosphate pesticide exposure and child IQ: The CHAMACOS cohort. NeuroToxicology, 56, 180–187. https://doi.org/10.1016/j.neuro.2016.07.01

Sharpton, T. J. 2014. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science, *5*. https://doi.org/10.3389/fpls.2014.00209

Shrestha, A., Kelty, C. A., Sivaganesan, M., Shanks, O. C., & Dorevitch, S. 2020. Fecal pollution source characterization at non-point source impacted beaches under dry and wet weather conditions. Water Research, 182, 116014. https://doi.org/10.1016/j.watres.2020.116014

Spietz, R. L., Williams, C. M., Rocap, G., & Horner-Devine, M. C. 2015. A dissolved oxygen threshold for shifts in bacterial community structure in a seasonally hypoxic estuary. PLOS ONE, 10(8). https://doi.org/10.1371/journal.pone.0135731 van der Stel, A.-X., & Wösten, M. M. 2019. Regulation of respiratory pathways in Campylobacterota: A Review. Frontiers in Microbiology, 10. https://doi.org/10.3389/fmicb.2019.01719

Taylor, E.M., Sweetkind, D.S., & Havens, J.C. 2017. Investigating the landscape of Arroyo Seco—Decoding the past—A teaching guide to climate-controlled landscape evolution in a tectonically active region. U.S. Geological Survey Circular, 1425, 44. https://doi.org/10.3133/c1425.

Thamdrup, B. 2012. New pathways and processes in the global nitrogen cycle. Annual Review of Ecology, Evolution, and Systematics, 43(1), 407–428. https://doi.org/10.1146/annurev-ecolsys-102710-145048

Thomas, P. D. 2016. The gene ontology and the meaning of biological function. Methods in Molecular Biology, pp. 15–24. https://doi.org/10.1007/978-1-4939-3743-1_2

Tomasek, A., Staley, C., Wang, P., Kaiser, T., Lurndahl, N., Kozarek, J. L., Hondzo, M., & Sadowsky, M. J. 2017. Increased denitrification rates associated with shifts in prokaryotic community composition caused by varying hydrologic connectivity. Frontiers in Microbiology, 8. https://doi.org/10.3389/fmicb.2017.02304

U.S. Customs and Border Protection. n.d. DRAFT: Water Quality Analysis Sampling Period January – June 2018. California Water Board. [Internet]. [cited 2021 Nov 23]; Available from https://www.waterboards.ca.gov/sandiego/board_info/agendas/2019/mar/item8/Item8_S D1_TijuanaTransboundarySummary.pdf

Venâncio, I., Luís, Â., Domingues, F., Oleastro, M., Pereira, L., & Ferreira, S. 2022. The prevalence of Arcobacteraceae in aquatic environments: A systematic review and meta-analysis. Pathogens, 11(2), 244. https://doi.org/10.3390/pathogens11020244

Vincent, S. G., Jennerjahn, T., & Ramasamy, K. 2021. Assessment of microbial structure and functions in coastal sediments. Microbial Communities in Coastal Sediments, 167–185. https://doi.org/10.1016/b978-0-12-815165-5.00006-6

Wade, W. 2002. Unculturable bacteria--the uncharacterized organisms that cause oral infections. JRSM, 95(2), 81–83. https://doi.org/10.1258/jrsm.95.2.81

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., Sievert, S. M., Simon, J., Campbell, B. J., Hanson, T. E., Woyke, T., Klotz, M. G., & Hugenholtz, P. 2017. Comparative genomic analysis of the class Epsilonproteobacteria and proposed reclassification to Epsilonbacteraeota (phyl. nov..). Frontiers in Microbiology, 8. https://doi.org/10.3389/fmicb.2017.00682

Wantzen, K. M., & Junk, W. J. 2006. Aquatic-terrestrial linkages from streams to rivers: Biotic hot spots and Hot moments. River Systems, 16(4), 595–611. https://doi.org/10.1127/lr/16/2006/595

Warnes, R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley., T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B. 2015. Gplots: Various R Programming Tools for Plotting Data.

Willis, A. D. 2019. Rarefaction, alpha diversity, and Statistics. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02407

Wooley, J. C., Godzik, A., & Friedberg, I. 2010. A primer on metagenomics. PLoS Computational Biology, *6*(2). https://doi.org/10.1371/journal.pcbi.1000667

Xiong, J., Liu, Y., Lin, X., Zhang, H., Zeng, J., Hou, J., Yang, Y., Yao, T., Knight, R., & Chu, H. 2012. Geographic distance and ph drive bacterial distribution in alkaline lake sediments across Tibetan Plateau. Environmental Microbiology, 14(9), 2457–2466. https://doi.org/10.1111/j.1462-2920.2012.02799.x

Yuan, X., Knelman, J. E., Gasarch, E., Wang, D., Nemergut, D. R., & Seastedt, T. R. 2016. Plant community and soil chemistry responses to long-term nitrogen inputs drive changes in alpine bacterial communities. Ecology, 97(6), 1543–1554. https://doi.org/10.1890/15-1160.1

Zeden, M. S., Burke, Ó., Vallely, M., Fingleton, C., & O'Gara, J. P. 2021. Exploring amino acid and peptide transporters as therapeutic targets to attenuate virulence and antibiotic resistance in Staphylococcus aureus. PLOS Pathogens, *17*(1). https://doi.org/10.1371/journal.ppat.1009093

Zhang, W.L., Li, Y., Wang, C., Wang, P.F., Hou, J., Yu, Z.J., Niu, L.H., & Wang, J. 2016. Modeling the biodegradation of bacterial community assembly linked antibiotics in river sediment using a deterministic–stochastic combined model. Environ. Sci. Technol. 50 (16), 8788–8798.

Zhu, Y., Weiss, E. C., Otto, M., Fey, P. D., Smeltzer, M. S., & Somerville, G. A. 2007. Staphylococcus aureus Biofilm Metabolism and the Influence of Arginine on Polysaccharide Intercellular Adhesin Synthesis, Biofilm Formation, and Pathogenesis. Infection and Immunity, *75*(9), 4219–4226. https://doi.org/10.1128/iai.00509-07

Scientific Protocols

Aline Biosciences

PCRClean DX
Retrieved from: https://alinebiosciences.com/wp-content/uploads/2017/03/Protocol-PCRClean-DX-v2.10_2-1.pdf

ThermoFisher

Qubit dsDNA HS Assay Kit
Retrieved from: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FQubit_dsDNA_HS_Assay_UG.pdf&title=VXNlciBHdWlkZTogUXViaXQgZHNETkEgSFMgQXNzYXkgS2l0cw==

Illumina

MiSeq System: Denature and Dilute Libraries
Retrieved from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-10.pdf

APPENDIX

Table A1. Nanodrop DNA quantification of samples for 16s sequencing.

Sample ID	Study Area	Sample Location	DNA Conc (ng/uL)	260/280
AS01	Salinas River Valley	Arroyo Seco	7.3	2.97
AS02	Salinas River Valley	Arroyo Seco	22.4	1.84
AS03	Salinas River Valley	Arroyo Seco	17.3	2.04
GC01	Mexican/US Border	Goat Canyon A	10	1.89
GC02	Mexican/US Border	Goat Canyon A	8.2	1.89
GC04	Mexican/US Border	Goat Canyon B	4.6	2.19
GC03	Mexican/US Border	Goat Canyon A	90.6	1.92
GC05	Mexican/US Border	Goat Canyon B	38.3	1.91
GC06	Mexican/US Border	Goat Canyon B	107.4	1.86
PL01	Mexican/US Border	Penasquitos Reserve	45.9	1.92
PL02	Mexican/US Border	Penasquitos Reserve	66.3	1.91
PL03	Mexican/US Border	Penasquitos Reserve	78.4	1.91
SRB01	Salinas River Valley	CDPR Salinas Site	61.8	1.98
SRB02	Salinas River Valley	CDPR Salinas Site	88.1	1.93
SRB03	Salinas River Valley	CDPR Salinas Site	154.7	1.91
SRA01	Salinas River Valley	Salinas River Refuge	109.6	1.95
SRA02	Salinas River Valley	Salinas River Refuge	76.9	1.91
SRA03	Salinas River Valley	Salinas River Refuge	186.1	1.91
SG01	Mexican/US Border	Smugglers Gulch	24.8	1.95
SG02	Mexican/US Border	Smugglers Gulch	25	1.92
SG03	Mexican/US Border	Smugglers Gulch	37.3	1.92
SB148	Bioreactor	Bioreactor19	103.4	1.65
SB160	Bioreactor	Bioreactor19	60.1	1.87
SB456	Bioreactor	Bioreactor19	146.7	1.82
SB460	Bioreactor	Bioreactor19	177.8	1.82
SB508	Bioreactor	Bioreactor20	71.7	1.93
SB516	Bioreactor	Bioreactor20	104.1	1.91
SB596	Bioreactor	Bioreactor20	29.5	1.93
SB604	Bioreactor	Bioreactor20	230.6	1.87
SB666	Bioreactor	Bioreactor21	34.8	2.02
SB670	Bioreactor	Bioreactor21	43.4	1.97
SB931	Bioreactor	Bioreactor21	31.6	2.1
SB935	Bioreactor	Bioreactor21	47.7	1.98

Table A2. Samples chosen to be sent in for whole metagenome shotgun sequencing.

Sample ID	Study Area	Sample Location	DNA Conc (ng/uL)	260/280	Notes
SB148	Bioreactor	Bioreactor19	103.4	1.65	Pre dosing
SB160	Bioreactor	Bioreactor19	60.1	1.87	Pre dosing
SB456	Bioreactor	Bioreactor19	146.7	1.82	After dosing
SB460	Bioreactor	Bioreactor19	177.8	1.82	After dosing
SB508	Bioreactor	Bioreactor20	71.7	1.93	Pre dosing
SB516	Bioreactor	Bioreactor20	104.1	1.91	Pre dosing
SB596	Bioreactor	Bioreactor20	27.4	2	After dosing
SB604	Bioreactor	Bioreactor20	230.6	1.87	After dosing
SB666	Bioreactor	Bioreactor21	34.8	2.02	Pre dosing
SB670	Bioreactor	Bioreactor21	43.4	1.97	Pre dosing
SB931	Bioreactor	Bioreactor21	31.6	2.1	After dosing
SB935	Bioreactor	Bioreactor21	47.7	1.98	After dosing
PL2	Mexican/US Border	Penasquitos Reserve	66.3	1.91	
PL3	Mexican/US Border	Penasquitos Reserve	78.4	1.91	
GC4	Mexican/US Border	Goat Canyon	90.6	1.92	
GC6	Mexican/US Border	Goat Canyon	107.4	1.86	
SG2	Mexican/US Border	Smuggler's Gulch	25	1.92	
SG3	Mexican/US Border	Smuggler's Gulch	37.3	1.92	
SRA1	Salinas River Valley	Salinas River Wildlife Refuge	109.6	1.95	
SRA3	Salinas River Valley	Salinas River Wildlife Refuge	186.1	1.91	
SRB2	Salinas River Valley	Salinas CDPR Site	88.1	1.93	
SRB3	Salinas River Valley	Salinas CDPR Site	154.7	1.91	
AS2	Salinas River Valley	Arroyo Seco	22.4	1.84	Failed Novogene QC
AS3	Salinas River Valley	Arroyo Seco	17.3	2.04	Failed Novogene QC

Table A3. DNA Qubit concentrations and corresponding volumes for final DNA pool.

Sample ID	Study Area	Sample Location	DNA Conc (ng/uL)	Volume to Add to Library Pool (uL)
PL1	Mexican/US Border	Penasquitos	0.682	26.00
PL2	Mexican/US Border	Penasquitos	0.389	26.00
PL3	Mexican/US Border	Penasquitos	0.664	26.00
GC1	Mexican/US Border	Goat Canyon	1.31	15.27
GC2	Mexican/US Border	Goat Canyon	Too Low	26.00
GC3	Mexican/US Border	Goat Canyon	1.07	18.69
GC4	Mexican/US Border	Goat Canyon	0.457	26.00
GC5	Mexican/US Border	Goat Canyon	0.443	26.00
GC6	Mexican/US Border	Goat Canyon	0.42	26.00
SG1	Mexican/US Border	Smuggler's Gulch	1.33	15.04
SG2	Mexican/US Border	Smuggler's Gulch	1.12	17.86
SG3	Mexican/US Border	Smuggler's Gulch	1.31	15.27
SRA1	Salinas River Valley	Salinas River Wildlife Refuge	1.55	12.90
SRA2	Salinas River Valley	Salinas River Wildlife Refuge	Too Low	26.00
SRA3	Salinas River Valley	Salinas River Wildlife Refuge	0.725	27.59
SRB1	Salinas River Valley	Salinas CDPR Site	2.54	7.87
SRB2	Salinas River Valley	Salinas CDPR Site	1.17	17.09
SRB3	Salinas River Valley	Salinas CDPR Site	0.99	20.20
AS1	Salinas River Valley	Arroyo Seco	2.98	6.71
AS2	Salinas River Valley	Arroyo Seco	1.26	15.87
AS3	Salinas River Valley	Arroyo Seco	0.938	21.32
SB148	Bioreactor	bioreactor19	1.83	10.93
SB160	Bioreactor	bioreactor19	2.09	9.57
SB456	Bioreactor	bioreactor19	3.87	5.17
SB460	Bioreactor	bioreactor19	0.764	26.00
SB508	Bioreactor	Bioreactor20	2.45	8.16
SB516	Bioreactor	Bioreactor20	1.66	12.05
SB596	Bioreactor	Bioreactor20	0.97	20.62
SB604	Bioreactor	Bioreactor20	1.86	10.75
SB666	Bioreactor	Bioreactor21	0.236	26.00
SB670	Bioreactor	Bioreactor21	Too Low	26.00
SB931	Bioreactor	Bioreactor21	0.118	26.00
SB935	Bioreactor	Bioreactor21	0.213	26.00
Control1	Negative Control	Negative Control	Too Low	26.00
Control2	Negative Control	Negative Control	0.536	26.00
Control3	Negative Control	Negative Control	1	26.00
Control4	Negative Control	Negative Control	0.268	26.00

Table A4. ANOVA covariate assumption 1 table.

	Df	**Sum Square**	**Mean Square**	**F value**	**P value**
Nitrate					
LocationSite	8	22113	2764.1	3.875	0.00601
Residuals	21	14980	713.3		
Phosphate					
LocationSite	8	372.7	46.59	983.4	<2e-16
Residuals	21	1	0.05		
Dissolved Oxygen					
LocationSite	8	17338	2167.3	457.2	<2e-16
Residuals	21	100	4.7		
pH					
LocationSite	8	7.076	0.8845	30.09	8.05E-10
Residuals	21	0.617	0.0294		
Temperature					
LocationSite	8	672.1	84.01	23.79	7.2E-09
Residuals	21	74.2	3.53		
Turbidity					
LocationSite	8	1834.3	229.29	37.84	9.06E-11
Residuals	21	127.3	6.06		
Salinity					
LocationSite	8	1259.8	157.5	519	<2e-16
Residuals	21	6.4	0.3		

Table A5. All environmental metadata

Sample ID	Location Site	Nitrate (mg/L)	Phosphate (mg/L)	pH	Temperature C	Salinity (sal)	Dissolved Oxygen (mg/L)	Turbidity (g/L)
SB670	Bioreactor21	167	0.83	7.41	16.389	1.74	0.29	0
SB666	Bioreactor21	152	0.8	7.37	16.37	1.78	0.08	10
SB516	Bioreactor20	112	1	7	17.7	2	5.4	0
SB508	Bioreactor20	90	1.3	6.97	17.8	3	2.1	0
SB596	Bioreactor20	64	0.28	7.21	17.94	1.59	0.62	9
SB160	Bioreactor19	59.4	1.14	7.4	17.6	0	7.5	11
SB604	Bioreactor20	57.59	0.3	7.19	19.1	1.57	0.2	6
SG01	Smugglers Gulch	41	12.5	8.4	18.54	0.25	8.39	0.329
SG02	Smugglers Gulch	41	12.5	8.4	18.54	0.25	8.39	0.329
SG03	Smugglers Gulch	41	12.5	8.4	18.54	0.25	8.39	0.329
SB460	Bioreactor19	36.2	1.11	7.5	26.8	3	1.4	14
SB456	Bioreactor19	33	1.08	7.5	19.8	2	1.9	17
GC01	Goat Canyon	30	4.7	8.81	21.15	0.52	5.26	0.679
GC02	Goat Canyon	30	4.7	8.81	21.15	0.52	5.26	0.679
GC03	Goat Canyon	30	4.7	8.81	21.15	0.52	5.26	0.679
GC04	Goat Canyon	28	5	8.18	20.4	0.6	0.35	0.778
GC05	Goat Canyon	28	5	8.18	20.4	0.6	0.35	0.778
GC06	Goat Canyon	28	5	8.18	20.4	0.6	0.35	0.778
SB935	Bioreactor21	27.69	0.61	7.36	19.427	1.55	1.17	2.69
SB148	Bioreactor19	17.4	1.29	7.8	15.8	1	10	22
SRB01	CDPR Salinas Site	15	2.9	7.81	9.74	0.83	84.8	1.066
SRB02	CDPR Salinas Site	15	2.9	7.81	9.74	0.83	84.8	1.066
SRB03	CDPR Salinas Site	15	2.9	7.81	9.74	0.83	84.8	1.066
AS01	Arroyo Seco	13	0.49	7.96	6.39	0.13	12.47	0.175
AS02	Arroyo Seco	13	0.49	7.96	6.39	0.13	12.47	0.175
AS03	Arroyo Seco	13	0.49	7.96	6.39	0.13	12.47	0.175
SRA01	Salinas River Refuge	11	2.5	7.69	11.69	22.51	6.98	23.22
SRA02	Salinas River Refuge	11	2.5	7.69	11.69	22.51	6.98	23.22
SRA03	Salinas River Refuge	11	2.5	7.69	11.69	22.51	6.98	23.22
SB931	Bioreactor21	10.83	0.44	7.32	18.64	1.55	1.02	1
PL01	Penasquitos Reserve	2.4	0.48	8.5	14.54	1.07	7.02	1.347
PL02	Penasquitos Reserve	2.4	0.48	8.5	14.54	1.07	7.02	1.347
PL03	Penasquitos Reserve	2.4	0.48	8.5	14.54	1.07	7.02	1.347

Table A6. Sequence counts of Level 3 GO categories within high-quality dereplicated bins found in corresponding samples.

Sample	transmembrane transport	biosynthetic process	cellular metabolic process	small molecule metabolic process	nitrogen compound metabolic process	organic substance metabolic process	primary metabolic process	establishment of localization	regulation of cellular process	cellular response to stimulus	cell communication	signal transduction	cellular component organization or biogenesis	regulation of metabolic process	Sequence total per sample
GC4	394	1116	1823	700	1618	1962	1789	490	313	0	0	0	0	0	10205
PL3	290	814	1330	545	1150	1378	1224	346	234	0	0	0	0	0	7311
SB148	143	458	746	283	635	664	736	170	128	0	0	0	0	0	3963
SB160	677	1938	3091	1302	2683	3239	2866	824	545	0	0	0	0	0	17165
SB456	907	3045	4965	1959	4367	5428	4900	1129	960	831	0	0	0	0	28491
SB460	0	211	323	118	307	346	311	0	0	0	0	0	0	0	1616
SB506	927	2589	4266	1529	3715	4404	3943	1147	923	765	0	0	0	0	24208
SB516	688	1732	3008	1077	2597	3270	2974	850	715	533	0	0	0	0	17444
SB596	380	1226	1949	709	1688	1973	1740	503	357	0	0	0	0	0	10525
SB604	178	412	727	252	580	707	606	225	151	125	0	0	116	0	4079
SB931	307	806	1367	458	1151	1415	1273	368	400	297	244	231	0	0	8317
SB935	0	37	60	15	59	63	60	7	14	0	0	0	0	11	326
SG2	354	1193	1850	701	1646	1897	1665	466	390	306	0	0	277	0	10745
SG3	148	417	673	248	607	691	602	198	107	0	0	0	0	0	3691
SRA1	200	723	1227	475	1050	1251	1115	287	0	0	0	0	0	0	6328
SRA3	260	974	1480	575	1325	1509	1323	352	0	0	0	0	227	0	8025

Table A7. Level 3 GO category terms that were overrepresented within Group A.

Overrepresented GO Terms				
GO Term	GO Name	GO Level 3 Group	Adj. P-value	P-value
GO:0003333	Amino Acid Transmembrane Transport	Transmembrane Transport	0.023338342	6.81E-04
GO:1903825	Organic Acid Transmembrane Transport	Transmembrane Transport	0.037285777	0.001305923
GO:1905039	Carboxylic Acid Transmembrane Transport	Transmembrane Transport	0.037285777	0.001305923
GO:1901566	Organonitrogen Compound Biosynthetic Process	Biosynthetic Process	1.49E-05	1.07E-07
GO:0008652	Amino Acid Biosynthetic Process	Biosynthetic Process	0.002604898	3.74E-05
GO:0043604	Amide Biosynthetic Process	Biosynthetic Process	0.011644557	3.03E-04
GO:0043043	Peptide Biosynthetic Process	Biosynthetic Process	0.02649818	7.97E-04

Table A8. Level 3 GO category terms that were underrepresented within Group A.

Underrepresented GO Terms				
GO Term	GO Name	GO Level 3 Group	Adj. P-value	P-value
GO:0015995	Chlorophyll Biosynthetic Process	Biosynthetic Process	4.49E-04	4.63E-06
GO:0009190	Cyclic Nucleotide Biosynthetic Process	Biosynthetic Process	4.49E-04	4.54E-06
GO:0009889	Regulation of Biosynthetic Process	Biosynthetic Process	0.002604898	3.86E-05
GO:0031326	Regulation of Cellular Biosynthetic Process	Biosynthetic Process	0.002699697	4.12E-05
GO:0010556	Regulation of Macromolecule Biosynthetic Process	Biosynthetic Process	0.003101809	5.43E-05
GO:2001141	Regulation of RNA Biosynthetic Process	Biosynthetic Process	0.003347523	6.16E-05
GO:0032774	RNA Biosynthetic Process	Biosynthetic Process	0.004985604	1.03E-04
GO:0042121	Alginic Acid Biosynthetic Process	Biosynthetic Process	0.020975691	5.88E-04
GO:0030494	Bacteriochlorophyll Biosynthetic Process	Biosynthetic Process	0.046470888	0.001711097
GO:0031323	Regulation of Cellular Metabolic Process	Cellular Metabolic Process	0.003504583	6.77E-05
GO:0051171	Regulation of Nitrogen Compound Metabolic Process	Nitrogen Compound Metabolic Process	0.002208303	2.97E-05
GO:0071704	Organic Substance Metabolic Process	Organic Substance Metabolic Process	0.002452314	3.41E-05
GO:0044238	Primary Metabolic Process	Primary Metabolic Process	2.61E-04	2.34E-06
GO:0080090	Regulation of Primary Metabolic Process	Primary Metabolic Process	0.0014085	1.71E-05
GO:0050794	Regulation of Cellular Process	Regulation of Cellular Process	1.42E-25	3.82E-28
GO:0051716	Cellular Response to Stimulus	Cellular Response to Stimulus	5.66E-25	1.78E-27
GO:0010646	Regulation of Cell Communication	Cell Communication	0.010028144	2.49E-04
GO:0007154	Cell Communication	Signal Transduction	3.06E-38	4.12E-41
GO:0019222	Regulation of Metabolic Process	Regulation of Metabolic Process	0.013465121	3.57E-04