STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

# Amplicon Sequencing Assessment to Measure Microbial Community Response from Heavy Metal Contaminated Soils in Savannah River Site, Tims Branch Watershed

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 20, 2019


Principal Investigators:

Juan Carlos Morales (DOE Fellow Student)
Florida International University

Pamela Weisenhorn, Mentor
DOE Office of Science, Argonne National Laboratory

Florida International University:

Ravi Gudavalli Ph.D. Program Manager
Leonel Lagos Ph.D., PMP® Program Director



Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0000598


FIU | Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

# ABSTRACT

**RESEARCH OPPORTUNITIES FOR TIMS BRANCH- STEED POND SYSTEM.**

Tims Branch watershed has received discharge from the SRS A/M Area containing heavy metals and radionuclides of which a large amount has attenuated in an abandoned farm pond predating former activities. An amplicon sequencing experiment was performed using Tims Branch watershed soils targeting the 16S rRNA gene to evaluate heavy metals (arsenic, cadmium, tin and nickel) and its effects on the microbial community diversity. The similarity criterion for OTU picking was set to 97 percent masking for inherent error in PCR and sequencing steps. Samples were compared using a parametric *t* test, ($p$ <0.05) between contaminated and control soils to determine the $\alpha$-diversity and significant bacterial genera among sample locations. It was found that 14 of 76 bacterial genera were significantly altered in low contamination soils (relative abundance ratio greater than 0.004%). In medium contaminated soils, 60 of 76 bacterial genera were significantly altered. Respectively, high contaminated soils revealed 62 of 76 significant bacterial taxa. The main phyla shared across all sample locations were *Proteobacteria, Acidobacteria, Chloroflexi* and *Verrucomicrobia.* Relative abundance comparison between groups reported significance in high contaminated soils. In addition, the negative effect of heavy metal loading on microbial activity was tested for changes in the microbial community present in the soil. These findings support the hypothesis that relative abundance and diversity is significantly altered in soils which are contaminated with heavy metals.

# TABLE OF CONTENTS

**ABSTRACT** ..... **iii**
**TABLE OF CONTENTS** ..... **iv**
**LIST OF FIGURES** ..... **v**
**LIST OF TABLES** ..... **vi**
**1. BACKGROUND** ..... **1**
**2. INTRODUCTION** ..... **3**
**3. EXECUTIVE SUMMARY** ..... **5**
**4. MATERIALS AND METHODS** ..... **6**
**5. RESULTS AND DISCUSSION** ..... **17**
**6. CONCLUSION** ..... Error! Bookmark not defined.
**7. REFERENCES** ..... **28**
**APPENDIX A.** ..... **30**

# LIST OF FIGURES

Figure 1.Tims Branch watershed (A/M area). The stream channels A-014, A-011, and Tims Branch are second order streams that discharge to the Upper Three Runs River and further flows into the Savannah River. ..... 1

Figure 2.The mean concentrations of the soil distributed across the Tims Branch stream. Sample locations match the sampling locations for the DNA extraction and 16S rRNA gene amplicon analysis (error bars = Standard error of the mean of four sample replicates per sample location). Sediment was collected from an average depth of (0.25-4.5 inches) from the riverbed. ..... 6

Figure 3. Soil types characteristics were identified and classified using Geographical Information Systems (GIS). The attributes for classification were downloaded from the USDA SSURGO database. ..... 7

Figure 4. Field sampling locations ranking according to the soil heavy metal analytical measurements from Tims branch watershed. ..... 8

Figure 5. Savannah River Site A/M Area along with outfalls and channel delineation ..... 9

Figure 6. Main workflow of the methods utilized in the preparation of the samples and data processing ..... 10

Figure 7. Materials and methods utilized in the identification of bacterial communities and their response to contaminated soils in Tims branch watershed. ..... 11

Figure 8. Flowchart of input files needed and created by QIIME pipeline used to analyze Tims Branch watershed soil samples. ..... 13

Figure 9. Forward sequence FASTQ file using the -i sequence_read_fps -n 2 < 10> first lines. This code helps the researcher identify the data and view the first 8 lines in the sequence along with the sample ID, raw sequence, accession run code and length. ..... 15

Figure 10. Sequence output for summary statistics using the summary. seqs python script. Summary table describing the sequence statistics from the biom file. ..... 17

Figure 11.Uses the sequences and assigns a taxonomic assignment based on default criteria rep_set_tax_assignment.txt ..... 18

Figure 12. Measured Operational Taxonomic Units (OTU's) of the relative abundance of Tims Branch soil samples bacterial taxa ..... 18

Figure 13. The relative abundances of major phyla detected in the Tims Branch watershed samples studied. Mean abundance values for each sample location over a map of the sampling locations at the Tims Branch Watershed (Aiken SC). Abundances of the mean are from 8 replicates. ..... 20

Figure 14.The samples collected expressed different taxonomic groups within each sample. The phylum (level 2) represents the percent abundance of taxa found across Tims Branch Watershed. The major detected taxa are Proteobacteria, Acidobacteria, Chloroflexi and *Verrucomicrobia*. (Tims Branch watershed, A/M area)- Abundances are the mean of 8 replicates. ..... 21

Figure 15. Image of the area sharing microbial taxa in 32 samples including 2 blanks collected in Tims Branch ..... 21

Figure 16. Distribution of relative abundance of bacterial taxa categorized according to sample location (S1-S4). From left to right (Blanks, Control- Above Beaver Pond 1, Rip Rap- Low contamination, Upstream Steeds Pond – Medium contamination, and Downstream of Steeds Pond- high contamination stream according to independent locations ..... 22

Figure 17. Rarefaction curves. The expressed graph was submitted to the Phylogenetic whole tree and Chao1 richness estimator. ..... 23

Figure 18. Alpha diversity metrics for each site location. The plot displays the median, upper and lower 25 % quartiles, minimum, maximum and outliers of α- diversity values. ..... 24

Figure 19. Unweighted PCoA plot identifying distance metrics of microbial communities. Samples close to each other represent close matching similarity with overlapping microbial communities in the phylogenetic trees. ..... 25

# LIST OF TABLES

Table 1. Distribution of soil heavy metal physiochemical concentrations ..... 7

Table 2. Descriptive table of sample collection design ..... 9

Table 3. Summary of CPU descriptive table used to run the samples ..... 12

Table 4. Essential input variables used to analyze 16S rRNA using QIIME ..... 14

Table 5. Strategies chosen to create an OTU table with the available data ..... 16

Table 6. Alpha diversity index was calculated for soils under different heavy metal treatments. 23

Table 7. Alpha diversity index (Chao1) comparison between groups using a two-sample t-test. 24

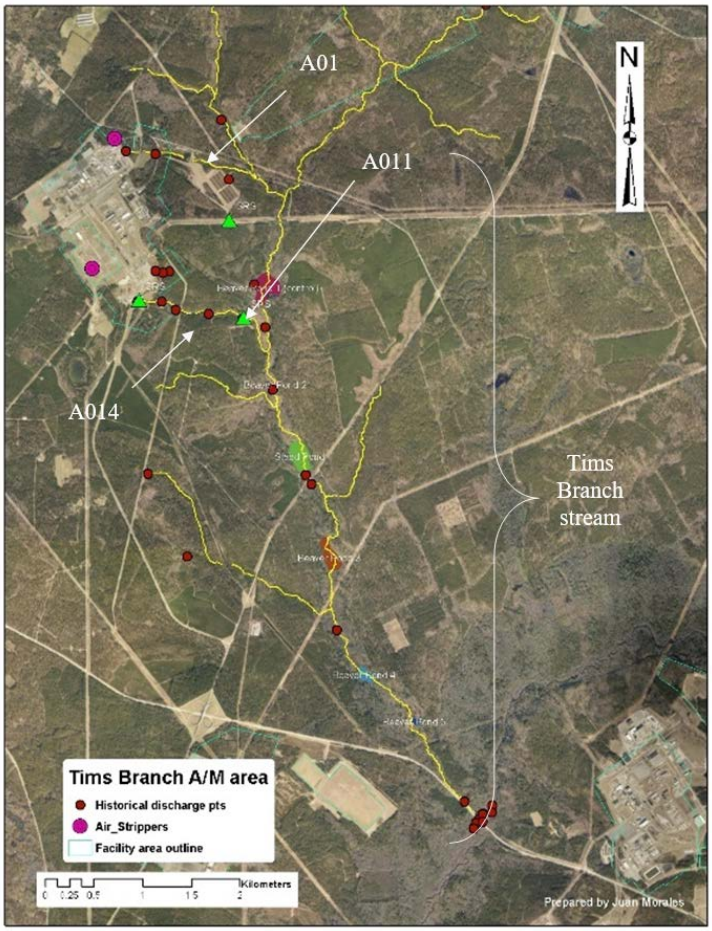
Table 8. Beta diversity metrics was determined by a two-way PERMANOVA analysis (Adonis function) from unweighted UniFrac distances matrices. ..... 26

Table 9. Sequencing mapping file used for downstream analysis ..... 30

Table 10. Summary of the sequences using split libraries to demultiplex the reads for all samples. ..... 31

# 1. BACKGROUND

**SAVANNAH RIVER SITE** is a former nuclear materials production site and is currently used as a research facility. It covers 777 $km^2$ bordering parts of Aiken, Barnwell and Allendale counties [1]. SRS is approximately 19 km south of Aiken, South Carolina, and 24 km southeast of Augusta, Georgia[2]. Tims Branch watershed (TBW) is located within SRS and is contained within the larger Upper Three Runs watershed, which is a sub-basin of the lower Savannah River Basin[3].




**Key Points:**

Tims Branch receives water from a treatment process facility that uses stannous chloride ($Sn_2Cl$) and air stripping to reduce the amount of mercury in the groundwater. Tims Branch waters originate from the facility outfalls in two areas – SRS National Laboratory to the north and the fueling target facility – M area to the west. The headwaters of Tims Branch intermittently falls into a losing stream (water falls below the water table and seeps into the ground).

**Figure 1. Tims Branch watershed (A/M area). The stream channels A-014, A-011, and Tims Branch are second order streams that discharge to the Upper Three Runs River and further flows into the Savannah River.**

In addition, Tims Branch stream is a small braided, marshy, second order stream that starts in the northern portion of SRS and passes through Beaver Ponds 1-5 and Steed Pond, and discharges into Upper Three Runs. The drainage area is nearly 16 $km^2$ with an average stream width

variation between 2 to 3 m [4]. Two other major tributaries, A-014 and A-011, both first order streams, are fed by groundwater pump stations located approximately 230 m apart [5].

Figure 1, identifies the research area and the historical discharges in which heavy metals and other contaminants were distributed as a result of the production of nuclear fuel. It is estimated that 43,500 kg of uranium (U) entered the Tims Branch system along with smaller quantities of aluminum (Al), nickel (Ni), copper (U), arsenic (As), chromium (Cr) and zinc (Zn) [1][5]. Currently, surface waters and sediments int TBW contain persistent levels of heavy metal concentrations (As, Cd, Cu, Ni, Pb, Hg, and U) and are suggested they can be detrimental to human and environmental health. Many complex human diseases have been correlated to the heavy metal distribution in surface waters leading to neurological diseases, cancers and organ toxicity [6]–[8].

**Remediation technologies implemented in TBW**

The U.S. Department of Energy's Office of Environmental Management (DOE-EM) in conjunction with the Applicable and Relevant or Appropriate Requirements (ARARs) found in the Record of Decisions (ROD) initiated remedial actions to reduce and limit mercury (Hg – species) concentrations in surface waters and methylmercury ($CH_3Hg^+$) levels which leads to biomagnification in aquatic life. It followed a treatment system for the removal of Hg – species in the northern headwaters of Tims Branch, mainly focused on the cessation of operations utilizing stannous chloride ($SnCl_2$) and air stripping as a source removal [12]. Respectively, this treatment system rapidly reduced the input of Hg – species into Tims Branch stream [13]. The reduction process was achieved by injecting $SnCl_2$ to reduce $Hg^{2+}$ to its elemental gaseous form – $Hg^0$ [14]. It was later confirmed by Looney et al., [16] that stoichiometric ratios greater than about 5 to 25, showed a relative complete removal with a final mercury ($Hg^{2+}$) concentration of <10 ng/L. Mercury concentrations were significantly decreased by groundwater treatment, however the concentration of inorganic metals present in the soils has not degraded.

# 2. INTRODUCTION

It is estimated that about 80 percent of the U and other heavy metals that were released remains in the Tims Branch system, of which 70 percent has been attenuated in an abandoned farm pond predating former activities. Since industrialization, however, runoff from the processing facilities has increased in the environment due to dynamic climatic changes in the environment. Despite great remediation progress in the area, Tims Branch watershed is an effective environmental sink for sequestering heavy metals in soils [9]. More, the legacy contamination effects in the soil often leads to disturbances in the microbial natural processes that are known to help decompose organic residues, form soil organic matter, and help with the mineralization process of nutrients [10]. Most of the soil microorganisms are in a geographic location often characterized as temperate region with metal oxide and organic rich soils contaminated with a complex mixture of metal species widely distributed along the stream.

There are however contradictory results regarding the impact of heavy metals on microbial species. Tipayno et al, (2018) suggests that the long term presence of heavy metal contamination in soils is correlated with changes in microbial community structures and can lead to the reduction of indigenous species [13]. In addition, culture independent methods identify that several species are reported to become tolerant in response to the chronic exposure to heavy metal toxicity varying across microbial communities. More, Azarbad et al., (2014) reports that several microbial operational taxonomic units (OTUs) are identified to survive and maintain functions in their communities when exposed to chronic exposure to heavy metals. More, he further identifies that the long-term functional levels of two distinct gradients of soil microbial communities were affected. Within the bacterial domain, *Actinobacteria,* are known into tolerating high concentrations of heavy metals, while the production of most soil bacteria in the same situation is limited.

There is increasing concern that Tims Branch watershed may eventually become a source for contaminants. The presence of certain microorganisms can affect metal species and mobility and thus playing an important role in the environmental fate and transport of metals and radionuclides.

Considering these threats, biomonitoring in the 21st century has opened a wide array of quantifying frameworks that enables researchers to describe and examine the impacts of environmental changes on ecosystem dynamics. The use of amplicon sequencing technology using the 16S rRNA gene has become a relatively easy way for comparative analysis for microbial community diversity, abundance and functional genes at greater sequencing depths from contaminated soils [14]. This method incorporates the use of ultra-deep sequencing of PCR products that efficiently check the variability of identification and characterization of microbial activity, function, diversity and evolution of soil microorganisms[12]. This technology is increasing due to its discovery rate of multiple species, a widely used method for the phylogeny and taxonomy studies, particularly in diverse samples.

- To highlight the importance of this study, we hypothesize that the relative abundance and diversity of species is significantly altered in soils which are contaminated with heavy metals in the Tims Branch watershed.

The research questions involving our analysis are as follows: Do heavy metals in soils, negatively decrease Operational Taxonomic Units (OTUs) when compared to other sample locations? Second, do the species richness decrease in contaminated samples? between samples? Lastly, are the α-alpha and β-beta diversity of species in low, mid and high contaminated soils significantly altered *p*- value < 0.05 compared to control samples?

**We plan to test our hypothesis by analyzing the bacterial community and diversity**

1. Hypothesis testing via *t*-test of α- diversity metrics
2. Hypothesis testing to determine how similar are the sample distances using PERMANOVA (Adonis)
3. Indicator species analysis (summarize taxa)

# 3. EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Development Initiative, an innovative program developed by the U.S. Department of Energy's Office of Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the summer of 2019, DOE Fellow intern Juan Morales spent 10 weeks doing a research summer internship at DOE-Office of Science, Argonne National Laboratory under the supervision and guidance of computational ecologist, Dr. Pamela Weisenhorn. The intern's project was initiated on June 3rd, 2019 and continued through August 10th, 2019 with the objectives to 1) use amplicon sequencing technology along with bioinformatics to compare bacterial communities between four sites in Tims Branch watershed. And, 2) to evaluate the percent relative abundance and diversity across the watershed.

# 4. MATERIALS AND METHODS

The sampling regions from various locations across the watershed were selected to obtain adequate mapping of the distribution of the contaminants. The sites were selected because of their known soil heavy metal concentration (As, Cd, Ni, and Sn) in which researchers from the Applied Research Center at FIU reported concentrations and soil type classification profiles during the summer of 2016 and 2017 [15]. Soil profile characteristics is represented in Figure 3 outlining the most dominant soils in Tims Branch watershed.

**Soil physiochemical characteristics**

Heavy metal concentrations recorded at these locations are shown in Table 1: S1-Control (As: 2.00 mg $kg^{-1}$, Cd: 2.00 mg $kg^{-1}$, Sn: 15.25 mg $kg^{-1}$ and Ni: 25.00 mg $kg^{-1}$), S2-Rip Rap (As: 7.86 mg $kg^{-1}$, Cd: 8.62 mg $kg^{-1}$, Sn: 183.5 mg $kg^{-1}$ and Ni: 25.00 mg $kg^{-1}$), S3-Upstream SP (As: 7.57 mg $kg^{-1}$, Cd: 8.45 mg $kg^{-1}$, Sn: 82.25 mg $kg^{-1}$ and Ni: 118.25 mg $kg^{-1}$) and S4- Downstream SP (As: 2.5 mg $kg^{-1}$, Cd: 6.25 mg $kg^{-1}$, 11.75 mg $kg^{-1}$, 25.00 mg $kg^{-1}$). Tims Branch aqueous concentrations for heavy metals are at or below regulatory limits - As: 10 mg $L^{-1}$, Cd: 0.005 mg $L^{-1}$, Sn: no data, indicating the effective sequestering of heavy metals in the river. A more descriptive representation of the soil physiochemical concentrations is described in Figure 2.

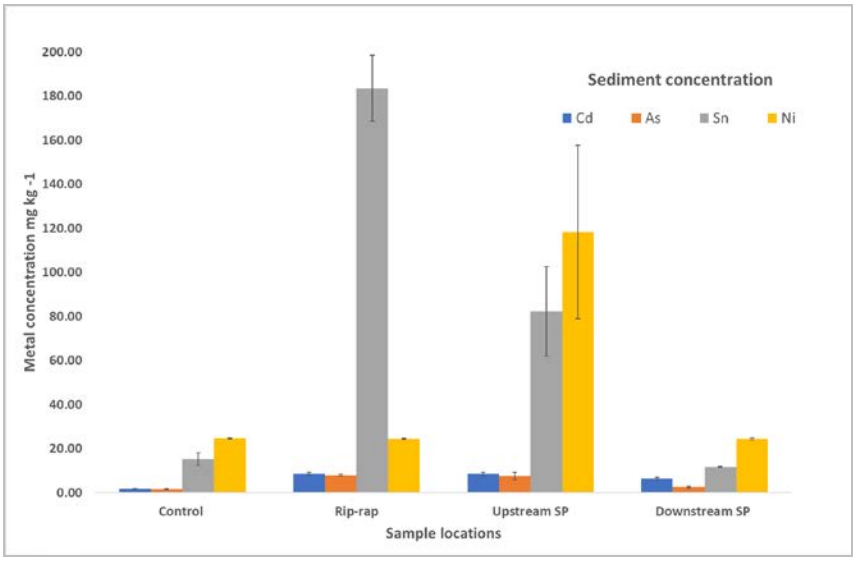



**Figure 2.The mean concentrations of the soil distributed across the Tims Branch stream. Sample locations match the sampling locations for the DNA extraction and 16S rRNA gene amplicon analysis (error bars = Standard Error of the Mean of four sample replicates per sample location). Sediment was collected from an average depth of (0.25-4.5 inches) from the riverbed.**

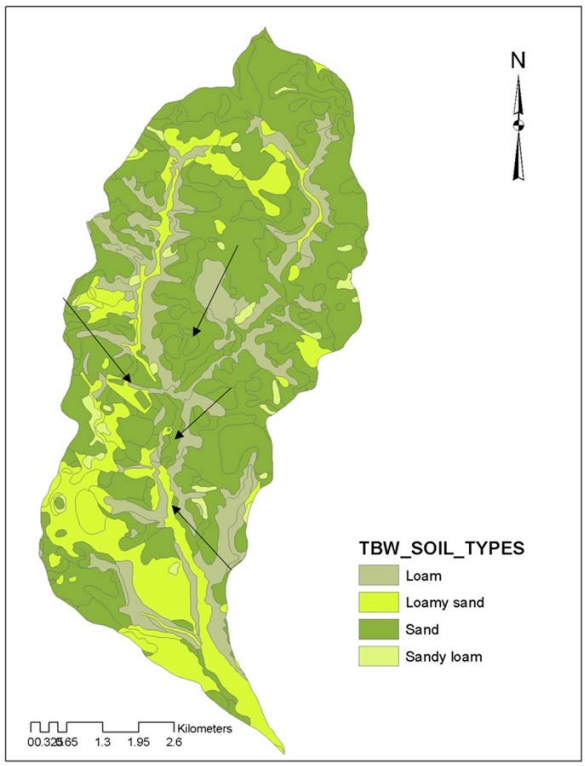



**Figure 3. Soil types characteristics were identified and classified using Geographical Information Systems (GIS). The attributes for classification were downloaded and classified according to soil types. Data used in this analysis was downloaded from the USDA SSURGO database.**

**Table 1. Distribution of soil heavy metal physiochemical concentrations**

| Soil samples (mg kg -1) | Total heavy metal concentration | | | | |
|---|---|---|---|---|---|
| | Texture | As | Cd | Sn | Ni |
| CN (Control) | Sand | 2.00 ± 0.1 | 7.86 ± 0.46 | 7.57 ± 1.72 | 2.5 ± 0.43 |
| CL (Downstream SP) | Sandy Loam | 2.00 ± 0.22 | 8.62 ± 0.53 | 8.45 ± 0.82 | 6.25 ± 0.65 |
| CH (Upstream SP) | Sand | 15.25 ± 2.81 | 183.5 ± 15.05 | 82.25 ± 20.38 | 11.75 ± 0.22 |
| CM (Rip rap) | Loam | 25 ± 0.22 | 25 ± 0.25 | 118.25 ± 39.41 | 25 ± 0.43 |

## FIELD SAMPLING DESIGN

All soil samples were collected during a single field sampling trip in Tims Branch watershed, Savannah River Site, Aiken, South Carolina **(33° 19' 2.172" N 81° 42' 54.288" W).** The collection sites are geographically isolated grounds in which the level of contamination was

ranked according to concentration status described in Figure 4. The highest contaminated area (CH region), medium contaminated (CM region) and lowest contaminated (CL region) soils were collected. Furthermore, a non-contaminated (CN region) which contains a similar soil parent material with the other selected regions, was chosen as the control.

| Control ** | Beaver Pond 1** |
|---|---|
| Low | Rip Rap / Met Lab channel |
| Mid | Upstream Steed's Pond |
| High | Downstream Steed's Pond |

**Figure 4. Field sampling locations ranking according to the soil heavy metal analytical measurements from Tims branch watershed.**

In total, eight random field-moist samples were taken from the top surface layer (0-6 cm) of the riverbed in each region and collected using sterile syringe corers with polythene bags for storage [1]. Each sample was an independent biological replicate with a total of thirty-two samples. The locations of each region are approximately 1 km apart and are identified Table 2. The collection design was initiated upstream of Beaver Pond 1, following the riprap and continuing downstream passing through Beaver pond (1-5) and Steed Pond, ending in lower Tims Branch. After sampling, soils were stored in an ice box at 4°C and shipped to the Sequencing Center at Argonne National Laboratory (ANL).

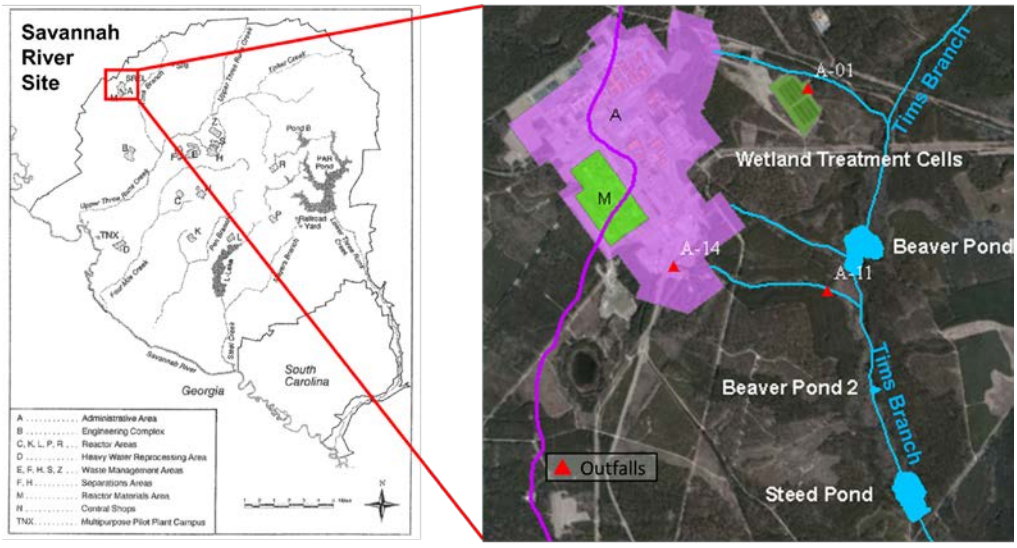



**Figure 5. Savannah River Site A/M Area along with outfalls and channel delineation**

**Table 2. Descriptive table of sample collection design**

| Sample ID | Ordinance of map Coordinates | | Elevation in meters | Number of samples (n) |
|---|---|---|---|---|
| S1- Beaver Pond 1 (control) | 33.31727 | -81.71508 | 83 | 8 |
| S2- Rip rap | 33.32485 | -81.71822 | 104 | 8 |
| S3- Upstream SP | 33.34035 | -81.7177 | 76 | 8 |
| S4- Downstream SP | 33.33175 | -81.72732 | 53 | 8 |

Total number of samples collected **(N = 32)**

Total number of sample replicates **(n=8)**

## EXPERIMENTAL SETUP

### Genomic DNA Extraction

Frozen soil was removed from the polythene bags, independently thawed and weighed. The subsamples of each soil replicate were manually loaded into the powertubes using the Qiagen DNEasy PowerSoil kit. The soil genomic DNA was extracted from 2g of soil, following the manufacturer's instructions[16]. Samples were separated into two batches and stored at -80°C. The total sample DNA was quantified using a Qubit fluorometer [3] and recorded. Figure 6, describes the workflow of the genomic soil DNA extraction.

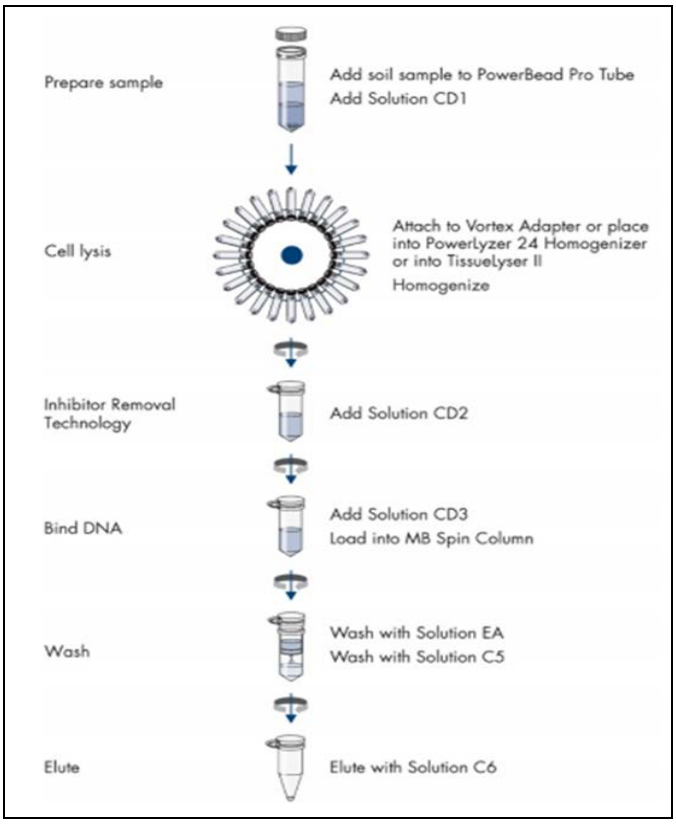



**Figure 6. Main workflow of the methods utilized in the preparation of the samples and data processing**

**Amplicon Library Preparation** Polymerase Chain Reaction (PCR) amplicon libraries were amplified targeting the 16S rRNA gene. We encoded the gene using primers 515 Forward (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806 Reverse (5′-GGACTACHVGGGTWTCTAAT-3′) for paired-end microbial community and were sequenced on the Illumina MiSeq platform at the Environmental Sample Preparation and Sequencing Facility (ESPSF) at Argonne National Laboratory (ANL) [16][17]. First, the genomic DNA was amplified for the region V4 (291 bp) of the 16S rRNA gene (515-806R) and tagged with specified sequencer adapters used for reference in the Illumina flowcell [18]. Each amplification reaction contained a 9.5 µM of MO BIO PCR water (Certified DNA-Free), 12.5 µL of QuantaBio's AccuStart II PCR ToughMix (2x concentration, 1x final), and 1 µL of template DNA.

The thermocycler conditions during the amplification cycles contained different denaturing phases: 94 °C for 3 minutes to denature the DNA, with 35 cycles at 94 °C for 45 s, 50 °C for 60 s, and 72 °C for 90 s; with a final extension of 10 min at 72 °C to ensure complete amplification. Amplicons were then quantified using PicoGreen (Invitrogen) and a plate reader (Infinite® 200 PRO, Tecan). Once quantified, amplified samples were pooled into a single tube so that each

amplicon is represented in equimolar amounts. The pool was then cleaned and quantified. After quantification, the molarity of the pool was determined and diluted down for sequencing purposes. Lastly, the amplicons were sequenced using Illumina MiSeq customized sequencing primers and procedures [19]. Figure 7, briefly describes the process of genomic extraction, bridge amplification cycle, sequencing of the DNA and visualization following the Illumina MiSeq methods.

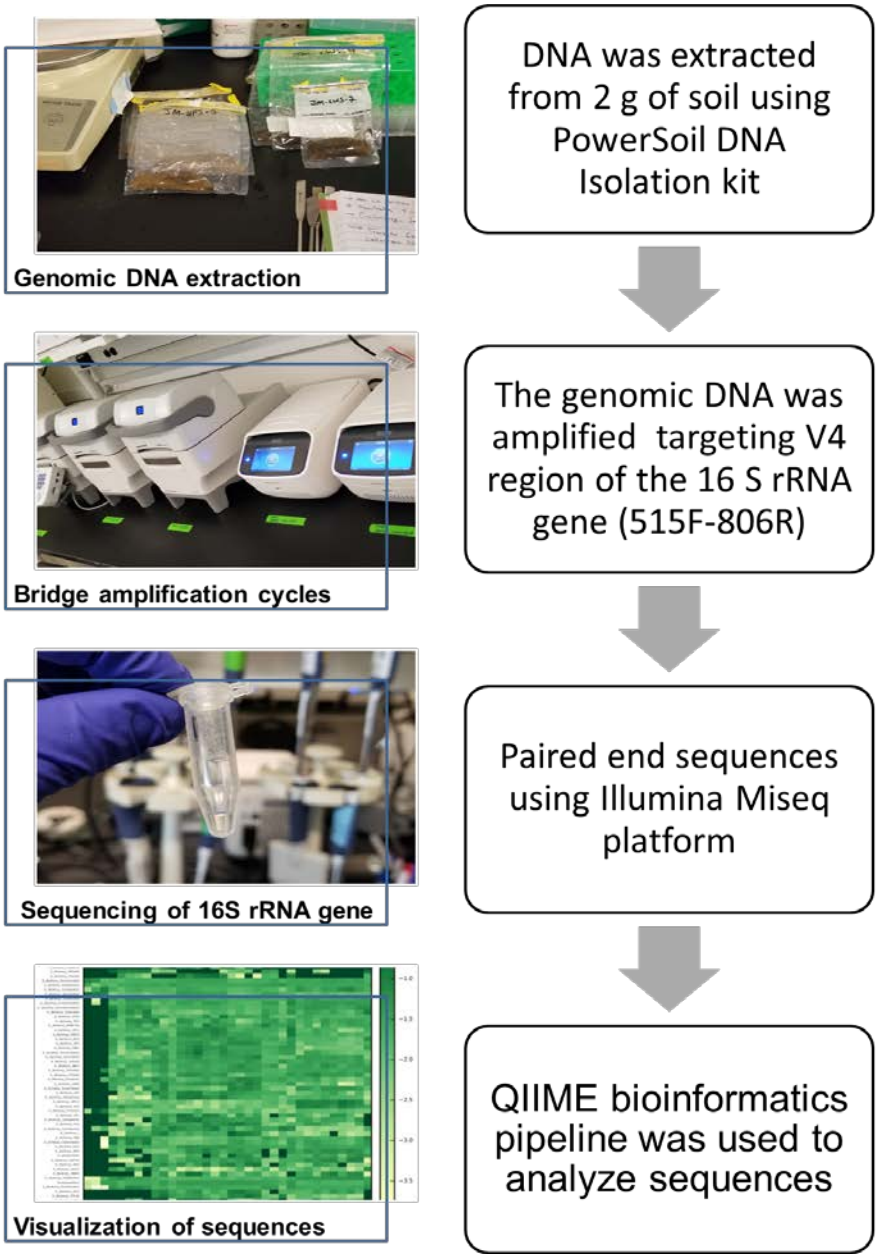



**Figure 7. Materials and methods utilized in the identification of bacterial communities and their response to contaminated soils in Tims branch watershed.**

## Data Analysis and Standardization

The Argonne National Laboratory Sequencing Center delivered the raw sequence reads via email to juan.morales@anl.gov and allowed for downloading using a link copied in the email. The account username and password to download the raw sequences was the same generated prior to entering the sequencing center; however, a link allowed for a ONE-TIME direct access to your data. Once the data was downloaded, the links paths becomes inaccessible. The data however, is still accessible via the website https://sequencing.bio.anl.gov and can be downloaded with the same user and password from before. The ANL Sequencing Center website includes the sequencing data (forward, reverse, barcodes and metadata). The raw sequences were then downloaded and copied to a local directory using a MacOS machine for downstream analysis. Table 3, describes in summary of the operating system and hardware used to analyze the raw sequence reads from ANL sequencing center.

**Table 3. Summary of CPU descriptive table used to run the samples**

| | |
|---|---|
| MacBook Pro | 13-inch, Mid 2012 |
| Processor | 2.5 GHz intel Core 5 |
| Memory | 8GB 1600 MHz DDR3 |
| Graphics | Intel HD Graphics 4000 1536MB |
| Serial Number | C02J8WGLDTY3 |

## Bioinformatics pipeline (QIIME)

The Quantitative Insights into Microbial Ecology software package (QIIME) is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. QIIME is designed to take users from raw sequencing data generated on the Illumina or other platforms through publication quality graphics and statistics. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations. QIIME has been applied to studies based on billions of sequences from tens of thousands of samples [20]. QIMMEs pipeline utilizes UNIX graphical interface which is viewed in the terminal window along with python wrapper scripts. Using QIIME allows the researcher to analyze microbial communities using a series of commands. The system can be run using Linux, MacOS or Windows operational system via a Virtual Box interface. QIIME has several limitations in which errors such as user defined syntax sensitivity can cause problems when executing the source code.

For this experiment, the MacOS was used which brings the Linux platform in the terminal window. Table 4, identifies the python-based scripts used for the processing of the FastQ files. More, QIIME pipeline enables the user to input datafiles called FASTQ, obtained from the ANL sequencing center. Once they are in the system, the FASTQ files will use the first four lines per sequence to identify and process the data. The flow chart of the input files needed to execute the program are describes in Figure 8, along with the directory path name.

*>User/juanmorales/desktop/MacQIIIME_1.9.1.-20150604_=S10.7/ANL_sequences*

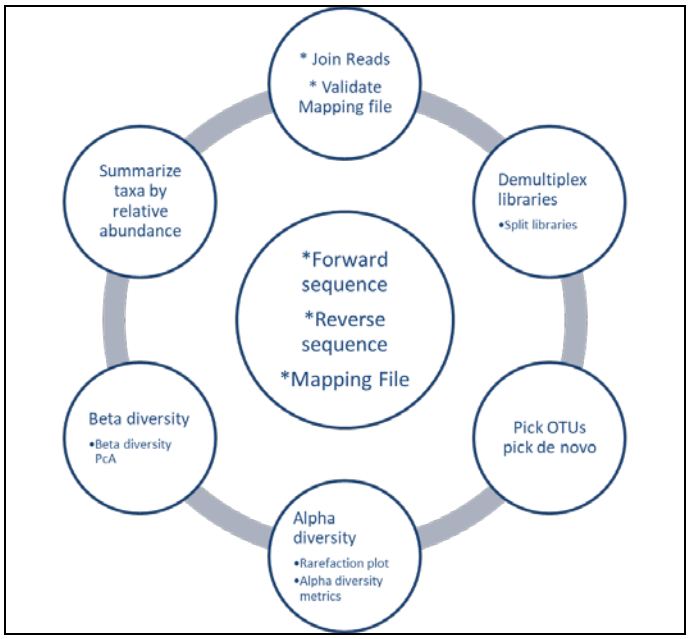



**Figure 8. Flowchart of input files needed and created by QIIME pipeline used to analyze Tims Branch watershed soil samples.**

QIIME uses a large community representative database to cross match representative taxa found in the samples. Open reference OTU strategy helps the user cluster sequences to a database which groups individual sequences against a platform, names Green genes, and then couples the sequences using de novo to the reference database. The default script used in this analysis was the *pick_de_novo_otus.py*. This workflow allows for the construction for the de novo OTU picking, taxonomy assignment, phylogenetic tree construction and OTU table construction.

**The output generated by the python script will contain the following:**

*(rep_set.tree)* → The phylogenetic tree describing the relationship of the raw sequence reads

*(otu_table.biom)* → The final OTU results

*(log_2019072252555004.tx)* → Basic summary of the run

*(uclust_picked_otus)* → parameter file clustering 97 percent of the sequences

*(uclust_assigned_taxonomy*) → taxonomic assignments stored in. biom table

*(pynast_aligned_seqs)* → Forward and reverse aligned sequences.

The first line in the sequence begins with a '@' character and is commonly followed by a sequence identifier and description. Line 2 is the raw sequence read letters or base pairs. Line 3 is a sequence that begins with a '+' sign and is followed by a second identifier. Line 4 encodes the quality values for the line 2 sequence and should have the same number of letters (symbols) as the letter in the sequence. Quality scores and mapping files should be corrected before demultiplexing the data.

**Table 4. Essential input variables used to analyze 16S rRNA using QIIME**

| QIIME Scripts | Input script files used for execution of command |
|---|---|
| Validate_mapping_file.py | -i mapping_file_corrected.txt |
| Join_paired_ends.py | -r reverse_sequence.fastq |
| Split_libraries_fastq.py | -b barcodes_file.fastq |
| Make_otu_heatmap.py | -i fastqjoin.join.fastq |
| Pick_de_novo_otus.py | -I fastqjoin.join.fastq |
| Summarize_taxa.py | -b fastqjoin.join_barcodes.fastq |
| Summarize_taxa_through_plots.py | -m Mapping_file_corrected.txt |
| | -o split_lib_TBW |
| | Additional input (--barcode_type12)- not all barcodes are the same |
| | -i out_table.biom |
| | -o heatmap.pdf |
| | -i seq.fna |
| | -o $PWD/uclust_otus |
| | -o taxa_Summary |
| | -I out_table.biom |
| | -o taxa_summary_TBW |
| | -i out_table.biom |
| | -m Mapping_file_corrected.txt |

**Mapping File**

The mapping file generated for input into QIIME was optimized according to the research question. Several parameters involving soil characteristics and concentration rankings were important categorical information to identify the samples along with the FASTQ raw files. The file generated was saved as a tab delimited .csv file. The mapping file included the Sample ID;

Barcode sequence – used for each sample; Linker Primer Sequence – used to amplify the sequence; Plate; well; description – sampling concentration variable and soil type and is described in Appendix, Table 9.

Moreover, any data generated was taken into consideration when assessing outliers. A preliminary test was used to account errors in the mapping file. The raw sequences were tested and validated with the mapping file matching the FASTQ raw data to ensure the mapping file does not have problems. The validate_mapping_file.txt corrects the file and unifies the data to match the sequences.

**Join sequence reads**

Default parameters were used for the 16S rRNA gene for the Illumina paired end reads. Using both forward and reverse amplicon sequence reads, the contigs were assembled using the *join_paired_ends.py*. The output file *(.fastq)* along with the *(fastqjoin.fastq.join_barcodes.fastq)* plus the respective mapping file *(Mapping_file_corrected.txt)* were analyzed and categorized. Figure 9, identifies the combined forward and reverse sequences using the Mapping_file_corrected.txt.

*>validate_mapping_file.py -m <mapping_fie path> -o Mapping_file_corrected.txt*

```
Juans-MacBook-Pro-2:ANL_sequences juanmorales$ head -n 10 forward_sequence.fastq
@M02149:300:000000000-CHFGF:1:1101:16559:1473 1:N:0:0
TCCTTCTTTTTCTCTCTTTCTTCTTCTTTCTTTNGCTTCCCTCTTTCTTAGGCTTTTTTTTCCTTCTTTTTTTCCCTCTCCTCTCTCCCCTTCCTTCT
TTCCTTTTCTCCTTCCTTTCTCTCTTTTTTTTTTTTTTCTTTTTCTTCCTCTT
+
>>>1>1B1311B1BBF1BB333D3133D3D3AD#000B0BFF001B221211110011///01DD111111>/>B@B000010@BB1BB?00001211
B1@11>2121BBF110122BB2211B111///////--0111=1/<=00000=
@M02149:300:000000000-CHFGF:1:1101:13541:1502 1:N:0:0
TCCTTTTTTTTCTCTCTTTTTTCTTCTTTCTTTNTCTTCCCTTTTTCTTCTTCTTTTTTTTCCTTCTTTTTTTCCCTCTCTTTTCTTCCCTTTTTTTT
TTCTTTTTCTCCTTTTTTTCTTTCTTTTTTTCTTTTTTTTTTTTCTTTCTTTT
```

**Figure 9. Forward sequence FASTQ file using the -i sequence_read_fps -n 2 < 10> first lines. This code helps the researcher identify the data and view the first 8 lines in the sequence along with the sample ID, raw sequence, accession run code and length.**

**Demultiplexing of sequences**

The sequences output file obtained from the previous analysis was demultiplexed using *(split_libraries_fastq.py).* This process was carefully executed because its importance in the identification process using the barcodes to identify which sequences belong to which after they

are all multiplexed together during the sequence run. Briefly, demultiplexed sequences output was renamed to *(histogram.txt), (seqs.fna)* and *(split_library_log.txt).*

**Operational taxonomic Units (OTUs) picking**

Continuing with the downstream analysis of the sequences, the Operational Taxonomic Units (OTUs - *de novo*) picking method for microbial analysis was used. This method allows the buildup of the OTU table using all sequences represented. There are other methods to analyze the data, however it was discovered that *de novo* clustering uses uncharacterized environments like soil and water, building the reference OTUs. Table 5, identifies all methods to analyze 16S rRNA gene with their advantages and disadvantages. The output file was used to pic a representative set of sequences using the pick_rep_set.py script and further assigned a taxonomy using *uclust* and *Green genes* database as the reference comparison database.

**Table 5. Strategies chosen to create an OTU table with the available data**

| **Picking OTUs Strategies** | **Advantages** | **Disadvantages** | **When to use—> depends on the data and the research question** |
|---|---|---|---|
| Open reference >*pick_open_reference_otus_py* | Clusters all sequences. Some work is parallelized, runs faster | Not parallelizable. Takes long | Uses any microbiome studies. Usually developers prefer this method. Advantages→ Clusters all sequences. Fast analysis. D isNot parallelizable, meaning it runs slow for big datasets. |
| Closed reference >*pick_closed_reference_otus.py* | Fast and parallelizable. Suitable for big datasets | Not possible to find new species | Data involving the human, mouse, skin and oral microbiome |
| De novo >*pick_de_novo_otus.py* | Clusters all sequences | Parallelizable is not enabled so slow for bid datasets | Data involving water, soil, environmental microbiome |

# 5. RESULTS AND DISCUSSION

**Soil Operational Taxonomic Unit.**

Data visualization makes big and small data easier for the human brain to understand. Good data visualization should place meaning into understanding the effects involved in pattern trends and outliers in the data. QIMME analyses each sequence displayed as count sequences and calculates the sequence length, mean and standard deviation. The *(.biom file*) represented in Figure 10 outlines a summary of 32 samples along with 2 blanks and their respective summary statistics. A total of 2.690.999 sequence counts highlights the amount of counts present in the sequences and 123.632 counts per sample symbolizes the maximum count per sample. Each observation was methodologically categorized into OTUs and the reference taxonomy can be identified in Figure 11. The assigned taxonomy for each OTU is representative upon the depth of sequence allowed by the confidence threshold of 80 percent, however it can be modified using the *(-c option script).*

```
Num samples: 34
Num observations: 212448
Total count: 2690999
Table density (fraction of non-zero values): 0.062

Counts/sample summary:
 Min: 720.0
 Max: 123632.0
 Median: 84389.500
 Mean: 79147.029
 Std. dev.: 26585.576
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 BLK-1: 720.0
 BLK-2: 13531.0
 JM-RR-3: 23213.0
 JM-UPS-8: 43665.0
 JM-UPS-4: 50258.0
 JM-CTRL-5: 61369.0
```

**Figure 10. Sequence output for summary statistics using the summary. seqs python script. Summary table describing the sequence statistics from the biom file.**

| | | | | |
|---|---|---|---|---|
| denovo15142 | k__Bacteria; p__Planctomycetes; c__Planctomycetia; o__Gemmatales; f__Gemmataceae; g__; s__ | 1.00 | 3 | |
| denovo123377 | Unassigned | 1.00 | 1 | |
| denovo123376 | k__Bacteria; p__Chloroflexi; c__Anaerolineae; o__A31; f__; g__; s__ | 1.00 | 3 | |
| denovo123375 | k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__BPC076; f__; g__; s__ | 1.00 | 3 | |
| denovo123374 | Unassigned | 1.00 | 1 | |
| denovo123373 | k__Bacteria; p__Acidobacteria; c__[Chloracidobacteria]; o__11-24; f__; g__; s__ | 1.00 | 3 | |
| denovo123372 | k__Bacteria; p__Acidobacteria; c__Acidobacteriia; o__Acidobacteriales; f__Koribacteraceae; g__Candidatus Koribacter; s__ | 1.00 | 3 | |
| denovo123371 | k__Bacteria; p__Acidobacteria; c__Solibacteres; o__Solibacterales; f__; g__; s__ | 0.67 | 3 | |
| denovo123370 | k__Bacteria; p__Acidobacteria; c__DA052; o__Ellin6513; f__; g__; s__ | 1.00 | 3 | |
| denovo123379 | k__Bacteria; p__Actinobacteria; c__Thermoleophilia; o__; f__; g__; s__ | 1.00 | 3 | |
| denovo123378 | k__Bacteria; p__NC10; c__12-24; o__JH-WHS47; f__; g__; s__ | 0.67 | 3 | |
| denovo41472 | k__Bacteria; p__Chloroflexi; c__Dehalococcoidetes; o__Dehalococcoidales; f__Dehalococcoidaceae; g__; s__ | 1.00 | 3 | |
| denovo41473 | k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Alteromonadales; f__211ds20; g__; s__ | 0.67 | 3 | |
| denovo41470 | k__Bacteria; p__Planctomycetes; c__Planctomycetia; o__Planctomycetales; f__Planctomycetaceae; g__Planctomyces; s__ | 1.00 | 3 | |
| denovo41471 | k__Bacteria; p__Actinobacteria; c__Thermoleophilia; o__Gaiellales; f__Gaiellaceae; g__; s__ | 1.00 | 3 | |
| denovo41476 | Unassigned | 1.00 | 1 | |
| denovo41477 | Unassigned | 1.00 | 1 | |
| denovo41474 | k__Bacteria; p__Acidobacteria; c__Acidobacteriia; o__Acidobacteriales; f__Koribacteraceae; g__; s__ | 0.67 | 3 | |
| denovo41475 | k__Bacteria; p__Verrucomicrobia; c__[Pedosphaerae]; o__[Pedosphaerales]; f__Ellin515; g__; s__ | 1.00 | 3 | |
| denovo41478 | k__Bacteria; p__OP3; c__kol11; o__; f__; g__; s__ | 1.00 | 3 | |
| denovo41479 | k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__ | 1.00 | 3 | |
| denovo15140 | k__Bacteria; p__Acidobacteria; c__Solibacteres; o__Solibacterales; f__Solibacteraceae; g__Candidatus Solibacter; s__ | 1.00 | 3 | |

**Figure 11.Uses the sequences and assigns a taxonomic assignment based on default criteria rep_set_tax_assignment.txt**

## Patterns of diversity (Heatmap of OTUs

Heatmap data analysis is an optimal application that uses color to interpret complete statistical data or trends. It uses a warm to cool color spectrum to visualize the data analytics, namely which parts of the data receive the most attention. Figure 12 heatmap for example represents the relative abundance of all taxa in all locations (0 being the least abundant to 4 highly abundant). *(Make_otu_heatmap.py)* script creates a heatmap of the sample's relative abundance. Each row corresponds to an individual OTU and each column corresponds to a single sample.

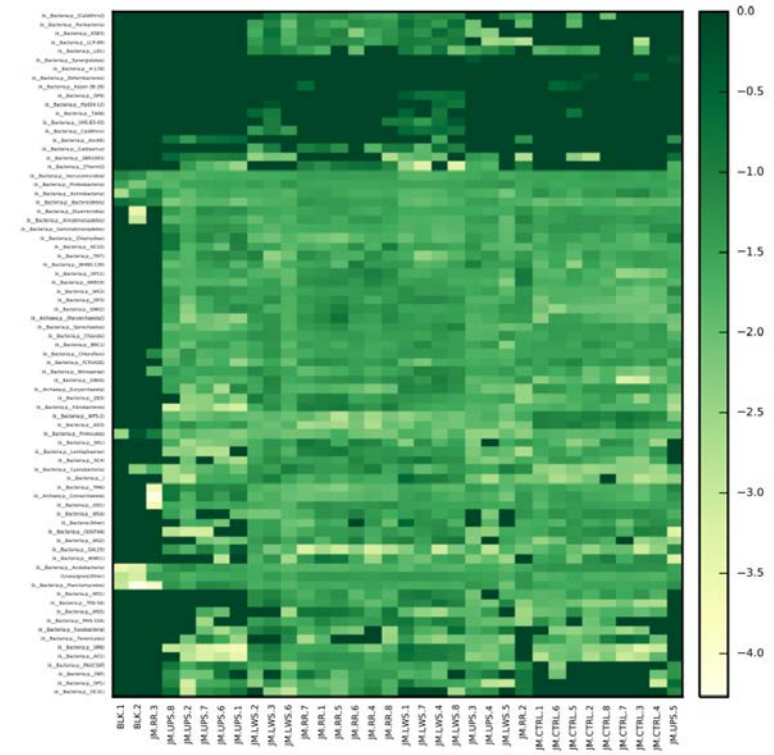



**Figure 12. Operational Taxonomic Units (OTU's) measuring the relative abundance of bacterial taxa across Tims Branch watershed soil samples.**

**Summary of each sample by taxonomic composition**

We summarized the samples by relative abundance to distinguish what microbial taxa is found in each sample community. The *summarize_taxa.py* and *plot_taxa_summary.py* algorithms were applied to identify the relative abundances by plotting the data. A text file was generated including the data per samples. We then analyzed and ranked the bacterial community based on their taxonomic group. Following ecology paradigms, it is important to associate bacterial communities of two or more different species occupying the same geographical area and time to account for a community or group of species.

**Microbial community composition**

A total of 14,512,909 16S rRNA gene sequences were grouped and analyzed. The collected 34 samples passed quality filtering steps explained in the Methods section. Figure 13, identifies 18 phyla collected and classified using a filter accounting for all > 1% of the total reads in this study. For control samples; Proteobacteria (27.25 %), Acidobacteria (27.08 %), Chloroflexi (8.51%), Verrucomicrobia (6.72%), Actinobacteria (3.83%), Planctomycetes (3.47%), Euryarchaeota (2.79%), Bacteroidetes (2.75%), Firmicutes (1.81%), Chlorobi (1.70%), Spirochaetes (1.40%), Crenarchaetes (1.32%), and Nitrospirae (1.04%). Rip rap samples; Proteobacteria (25.34%), Bacteroidetes (10.77%), Chloroflexi (8.72%), Verrucomicrobia (7.85%), Acidobacteria (7.36%), Cyanobacteria (4.58%), Planctomycetes (3.00%), Spirochaetes (2.85%), Nitrospirae (2.53%), Firmicutes (2.14%), Chlorobi (2.04%), Actinobacteria (1.66%), OP3 (1.31%), and Fibrobacteres (1.09%). Upstream Steeds Pond samples; Proteobacteria (37.04 %), Acidobacteria (11.64%), Bacteroidetes (8.95%), Planctomycetes (6.29%), Verrucomicrobia (4.98%), Chloroflexi (4.71%), Nitrospirae (3.48%), Cyanobacteria (3.46%), Crenarchaeota (1.89%), Chlamydiae (1.37%), and Actinobacteria (1.22%). Lastly, the downstream Steeds Pond samples; Proteobacteria (25.34%), Bacteroidetes (10.77%), Chloroflexi (8.72%), Verrucomicrobia (7.85%), Acidobacteria (7.36%), Cyanobacteria (4.58%), Planctomycetes (3.00%), Spirochaetes (2.85%), Nitrospirae (2.53%), Firmicutes (2.14%), Chlorobi (2.04%), Actinobacteria (1.66%), OP3 (1.31%) and Fibrobacteres (1.09%). The area utilized by the different taxa across treatment groups and control is represented in Figure 15. As important, Figure 16 illustrates the distribution of relative abundance of bacterial taxa categorized according to sample location (S1-S4). From left to right (Blanks, Control- Above Beaver Pond 1, Rip Rap-

Low contamination, Upstream Steeds Pond – Medium contamination, and Downstream of Steeds Pond- high contamination stream according to independent locations.

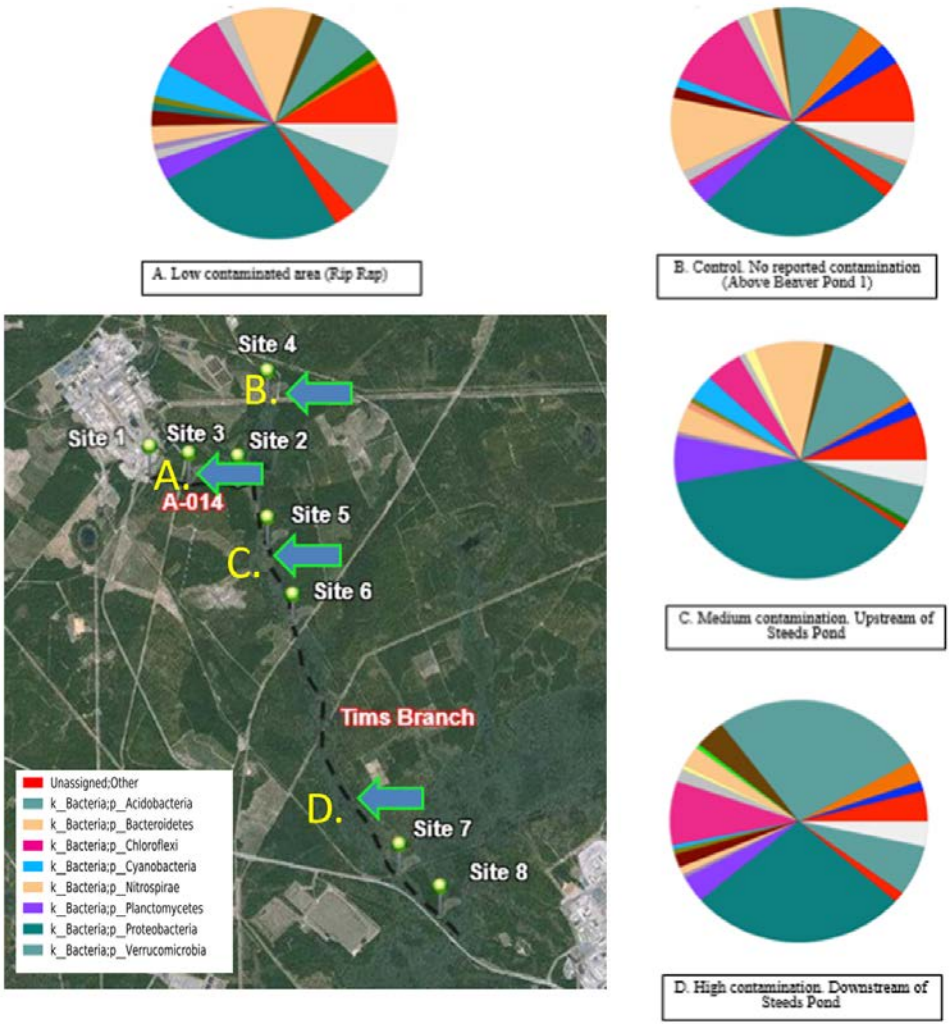



**Figure 13. The relative abundances of major phyla detected in the Tims Branch watershed samples studied. Mean abundance values for each sample location over a map of the sampling locations at the Tims Branch Watershed (Aiken SC). Abundances of the mean are from 8 replicates.**

Species ranking classification were also analyzed using the means and standard errors of the mean to account for the relative abundance of the most dominant species present in TBW. Outlined in Figure 14 the species ranking levels can be optimized by changing the scritpt to account for different parameters such as the the -L paramenter in the python script (1= Kingdom, 2 = Phylum, 3 = Class, 4 = Order, 5= Family, 6= Genus, and 7= Species.

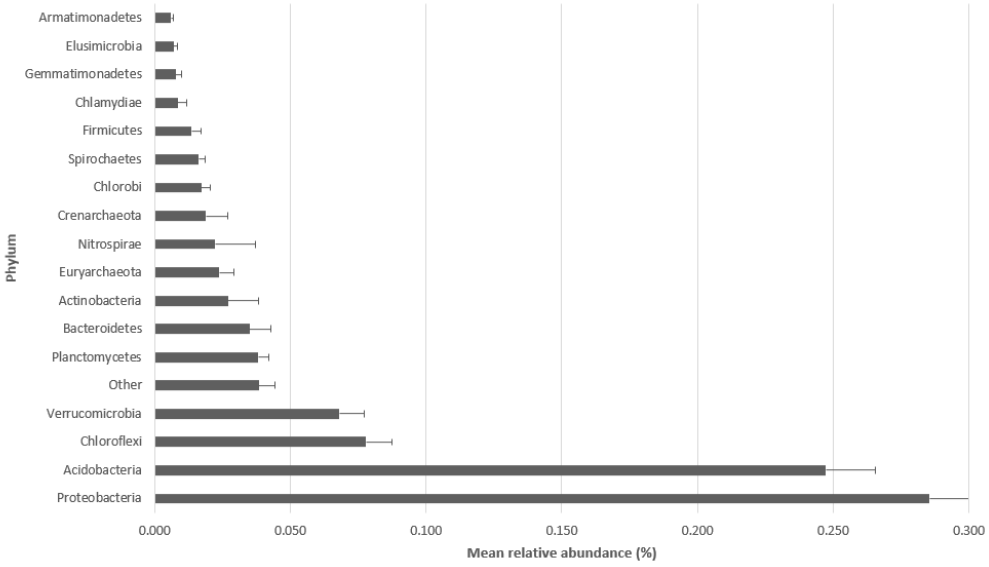



**Figure 14. The samples collected expressed different taxonomic groups within each sample. The phylum (level 2) represents the percent abundance of taxa found across Tims Branch Watershed. The major detected taxa are *Proteobacteria, Acidobacteria, Chloroflexi and Verrucomicrobia*. (Tims Branch watershed, A/M area)- Abundances are the mean of 8 replicates.**

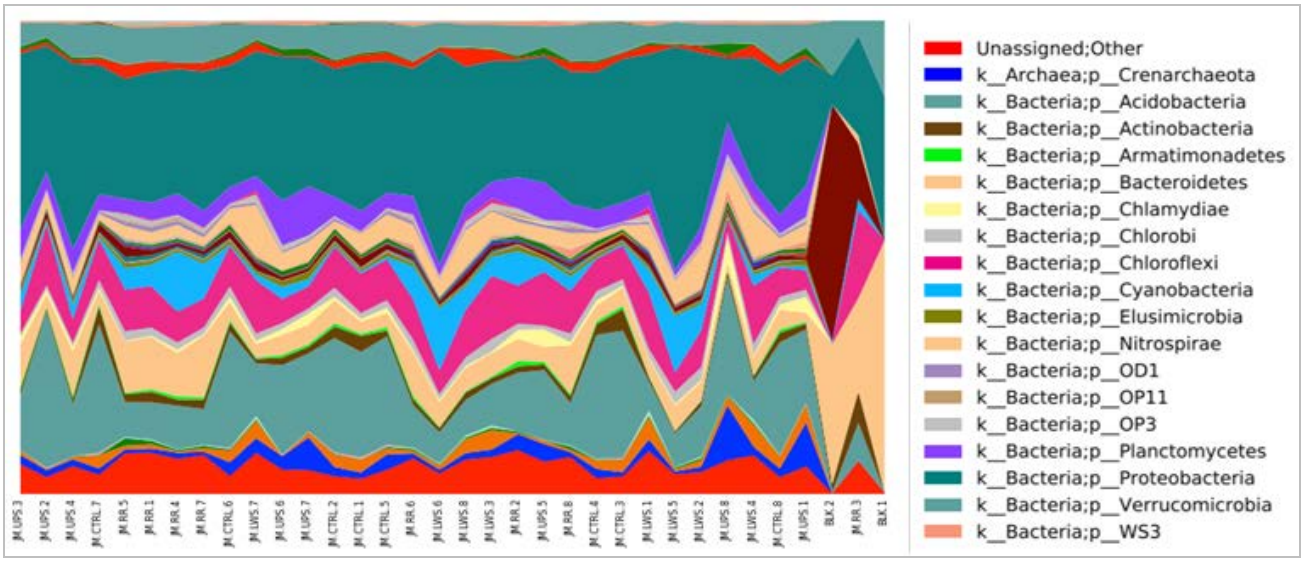



**Figure 15. Image of the area sharing microbial taxa in 32 samples including 2 blanks collected in Tims Branch**

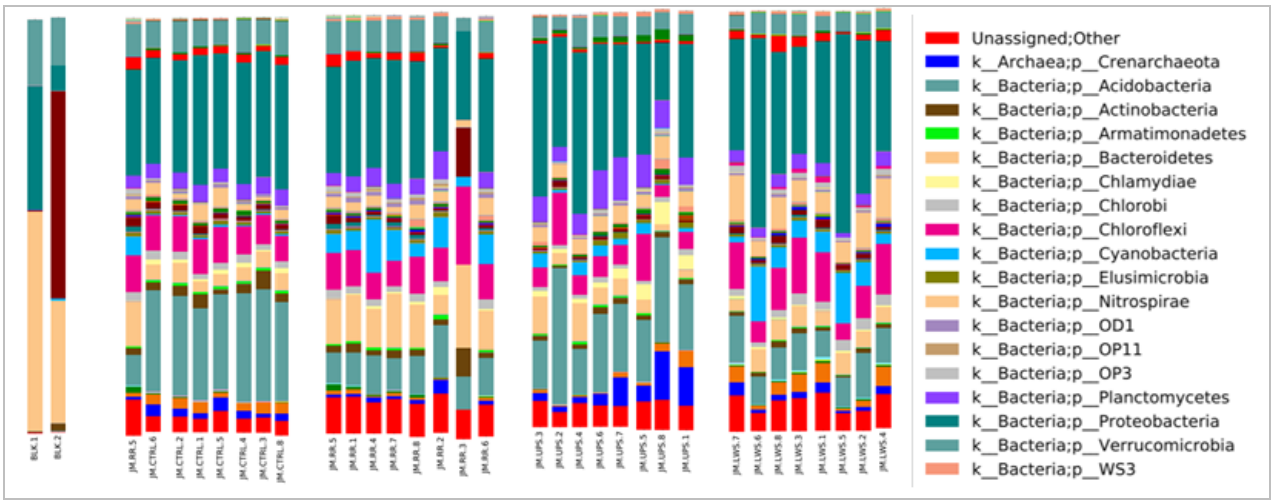



**Figure 16. Distribution of relative abundance of bacterial taxa categorized according to sample location (S1-S4). From left to right (Blanks, Control- Above Beaver Pond 1, Rip Rap- Low contamination, Upstream Steeds Pond – Medium contamination, and Downstream of Steeds Pond- high contamination stream according to independent locations**

### Rarefaction curve estimators

We calculated the alpha diversity metrics by using different metrics listed below: The phylogenetic diversity whole tree (*PD_whole_tree*) and and *Chao1*. **Error! Reference source not found.**, illustrates 3 units of measurement in which the amount of *observed species*, *Shannon index* and *Chao1* richness indicators are compared amongst groups. This type of comparative analysis can help answer questions such as how many species are there present in each sample location.

Research suggests that alpha diversity will increase within greater sequencing depth amount. More, the rarefaction curves enable us to compare the alpha diversity values versus the number included in the sequences. *Chao1* was used in this study and we are computing and predicting the OTUs richness compared to low sequence reads all the way to a high depth sequencing. After running the rarefaction script, we can observe the curve for *Chao1* with an html pipeline categorizing for soil types as the legend. The goal of this algorithm is to compute the sequencing depth which affects both alpha and beta diversity related experiments. Rarefaction curves identified in Figure 17 uses two different metrics to estimate the amount of sequences expressed per sample group. A. *Chao1* richness estimator. B. *PD_whole tree* or phylogenetic ehole tree richness estimator. Both metrics use a 97 % saturation in which richness and diversity index are illustrated.

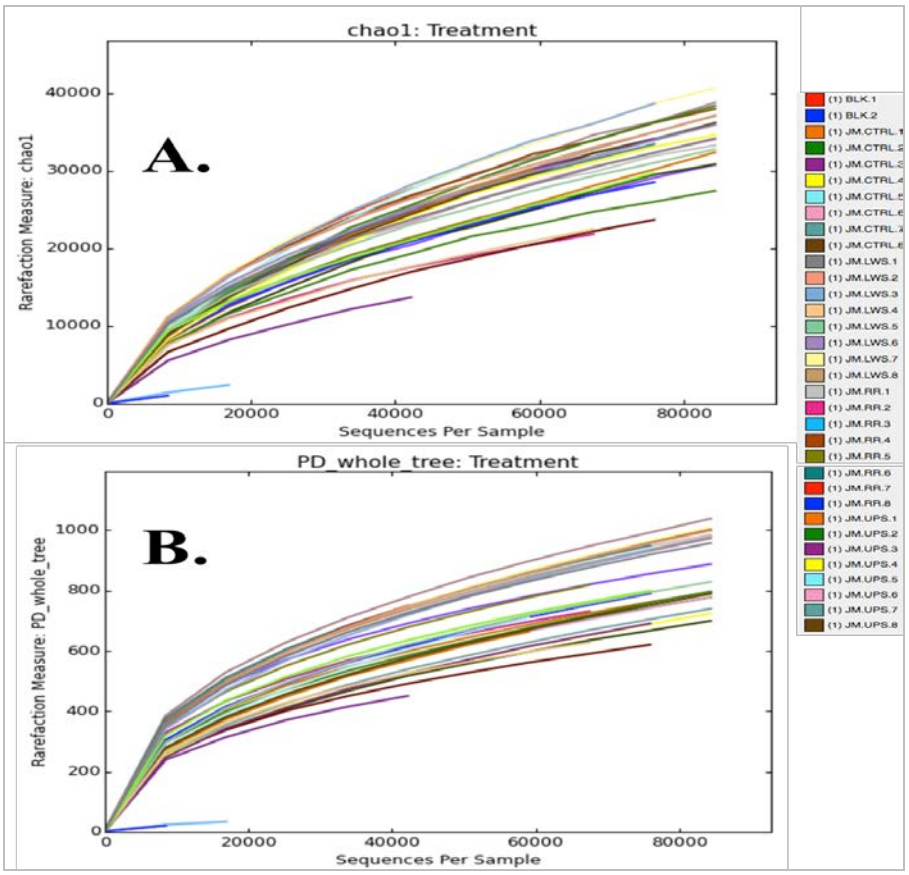



**Figure 17. Rarefaction curves identifying two different metrics used to estimate the amount of sequences expressed per sample group. A. Chao1 richness estimator. B. PD_whole tree or phylogenetic ehole tree richness estimator. Both metrics use a 97 % saturation in which richness and diversity index are illustrated.**

**Table 6. Alpha diversity index was calculated for soils under different heavy metal treatments.**

| Heavy metal treatment | Observed species OTUs | Shannon Index | Chao1 Index |
|---|---|---|---|
| Control (S1) | 86.5 ± 2.87 | 6.32 ± 0.07 | 563.72 ± 200.54 |
| Low (S2) | 87.75 ± 18.58 | 6.33 ± 0.50 | 1146.40 ± 620.71 |
| Mid (S3) | 89.25 ± 7.28 | 6.39 ± 0.18 | 885.84 ± 646.15 |
| High (S4) | 88.87 ± 6.12 | 6.35 ± 0.20 | 906.64 ± 631.82 |

Each value is a mean ± SD (n=8 in each group).
S1-Control (As: 2.00 mg $kg^{-1}$, Cd: 2.00 mg $kg^{-1}$, Sn: 15.25 mg $kg^{-1}$ and Ni: 25.00 mg $kg^{-1}$), S2-Rip Rap (As: 7.86 mg $kg^{-1}$, Cd: 8.62 mg $kg^{-1}$, Sn: 183.5 mg $kg^{-1}$ and Ni: 25.00 mg $kg^{-1}$), S3-Upstream SP (As: 7.57 mg $kg^{-1}$, Cd: 8.45 mg $kg^{-1}$, Sn: 82.25 mg $kg^{-1}$ and Ni: 118.25 mg $kg^{-1}$) and S4-Downstream SP (As: 2.5 mg $kg^{-1}$, Cd: 6.25 mg $kg^{-1}$, 11.75 mg $kg^{-1}$, 25.00 mg $kg^{-1}$).

## Statistical analysis

QIIME version 1.9.1 using the UNIX interface was used to perform comparison among groups. The *group_significance.py* script allowed for the comparison amongst OTU frequencies across each sample groups. Our analysis determined the significant differences between the OTU abundance in the different sample groups. The sample grouping was determined by the *-c option* in the terminal interface. Here suggest that there is a statistical significance among control and high treatment group and quantitatively speaking is illustrated in Table 7. Below, Figure 18 illustrates the comparison among control versus treatment groups using a two-tailed *t*-test testing for significance with a criterion of *p* value < 0.05. The plot displays the median, upper and lower 25 % quartiles, minimum, maximum and outliers of α- diversity values.

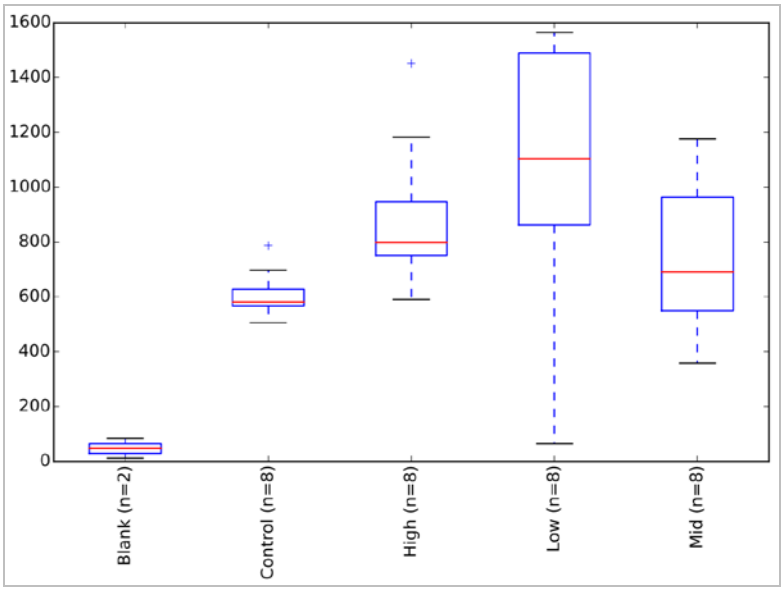



**Figure 18. Alpha diversity metrics for each site location. The plot displays the median, upper and lower 25 % quartiles, minimum, maximum and outliers of α- diversity values.**

**Table 7. Alpha diversity index (Chao1) comparison between groups using a two-sample t-test**

| Contaminated soils vs control | *p*-value |
|---|---|
| Low | 0.36 |
| Medium | 1.0 |
| High | **0.02**** |

n=8 in each group
Values are statistically significant at p<0.05**. All sample groups (Rip-rap low, Upstream SP Mid, and Downstream SP-High) are individually compared with normal control samples.

***Beta diversity metrics***

In order to determine the trends of differences and similarities between samples, Principal Coordinates Analysis (PCoA) was used. Out of the many metrics used to test distances between treatment versus control group samples this technique is among the best technique amongst microbial ecologists. QIIME uses EMPeror, a next generation tool that computes, visualizes and interprets high throughput microbial ecology datasets[21]. To test whether beta diversity is consistent with the sequencing technology the unweighted forms of analysis was computed using UniFrac metrics. Principal coordinate matrices (unweighted) was computed and three subfolders for each distance metric and 3D PCoA plots were generated. In addition, we analyzed the strength and statistical significance of sample groupings using a distance matrix as the primary input. R's vegan and ape packages were used to compute many of these methods, and for the ones that are not, their implementations are based on the implementations found in those packages. We performed a PERMANOVA analysis and significance was determined via a two-way ANOVA identified in Table 8. Unweight UniFRAC distances in Figure 19 revealed close similarities in phylogenetic diversities among treatments conditions.

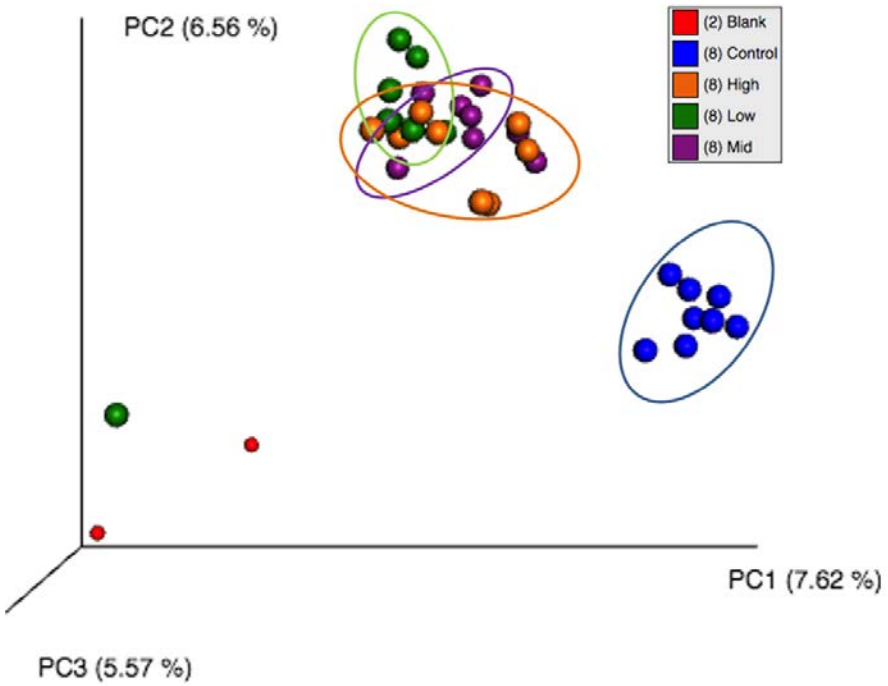



**Figure 19. Unweighted PCoA plot identifying distance metrics of microbial communities. Samples close to each other represent close matching similarity with overlapping microbial communities in the phylogenetic trees.**

**Table 8. Beta diversity metrics was determined by a two-way PERMANOVA analysis (Adonis function) from unweighted UniFrac distances matrices.**

| | SS | DF | MS | F | R2 | *p*-value |
|---|---|---|---|---|---|---|
| Group distance | 2.49 | 4 | 0.62 | 1.971 | 0.214 | **0.001**** |
| Residuals | 9.16 | 29 | 0.31 | | | |
| Total | 11.65 | 33 | | | | |

Values are statistically significant at p<0.05**. All sample distances between distances (Rip-rap low, Upstream SP Mid, and Downstream SP-High) are individually compared with control samples.

The

# 6. CONCLUSION

Here, we evaluate recent advances in nucleic acid extraction procedures and Next Generation Sequencing (NGS) technology to study and compare microbial communities in relationship to their environment. Consequently, the analysis led us to identify taxon and phylogenetic approaches using bioinformatics strategies. We determined that there are two groups from which we can make comparisons. The comparisons can be categorized into two groups: The within sample diversity (α- diversity) and the between sample diversity (β- diversity) comparison approaches. We determined that the sequence depths can influence the alpha and beta diversity metrics as shown by the rarefaction curves. More, the taxonomic classification was also evaluated using random selection of subsets of sequences in each sample, ensuring the sample number of reads are equivalent to those in the smallest sample.

We determined that all bacterial communities exposed to different heavy metal concentrations were dominated by four major groups (*Proteobacteria, Acidobacteria, Chloroflexi* and V*errucomicrobia).* The dominant phyla *Proteobacteria, Acidobacteria and Chloroflexi* accounted for 61 % of the relative abundance. Soil bacterial α- diversity, expressed as observed species richness metric and Shannon`s diversity index, was the highest in sample location four (high contamination).The lowest observed was at the control location. In additon, species richness increased when concentrations of heavy metals were localized, inferring that the metals can improve microbial community structure. Our distance analysis using PERMANOVA analysis between contaminated groups tested significant using the Adonis {F (4,29) =1.9712; $p < 0.001$**}. Finally, the PERMANOVA ($R^2$) identified a 20 percent variation explained between groups ($R^2$ =0.214)

This analysis proves to be beneficial in detecting microbial communities altered by contaminated soils. Moreover, certain concentrations of heavy metals can improve microbial community structures.

# 7. REFERENCES


1. Edwards PG, Gaines KF, Bryan AL, Novak JM, Blas SA. Are U, Ni, and Hg an Environmental Risk within a RCRA/CERCLA Unit on the U.S. Department of Energy's Savannah River Site? Hum Ecol Risk Assess. 2014;20(6):1565–89.
2. U.S. Department of Energy. Environmental Impact Statement. 1996 [cited 2019 Mar 19]; Available from: https://www.srs.gov/general/pubs/envbul/documents/EIS-0218-FEIS-v02-1996_0.pdf
3. Merem EC, Wesley J, Nwagboso E, Fageir S, Nichols S. Analyzing Environmental Issues in the Lower Savannah Watershed , in Georgia and South Carolina. 2015;5(1):1–20.
4. Batson VL, Bertsch PM, Herbert BE. Transport of anthropogenic uranium from sediments to surface waters during episodic storm events. J Environ Qual. 1996;25(5):1129–37.
5. Savannah River National Laboratory. Sediment Transport Studies in Tims Branch [Internet]. 1986 [cited 2018 Sep 28]. Available from: https://www.osti.gov/servlets/purl/6860203
6. Cannon JR, Greenamyre JT. The role of environmental exposures in neurodegeneration and neurodegenerative diseases. Toxicol Sci. 2011;124(2):225–50.
7. Vastrad B, Vastrad C, Tengli A, Iliger S. Identification of differentially expressed genes regulated by molecular signature in breast cancer-associated fibroblasts by bioinformatics analysis. Arch Gynecol Obstet. 2018;297(1):161–83.
8. Lee J, Freeman JL. Zebrafish as a model for investigating developmental lead (Pb) neurotoxicity as a risk factor in adult neurodegenerative disease: A mini-review. 2014 [cited 2018 Aug 30]; Available from: http://dx.doi.org/10.1016/j.neuro.2014.03.008
9. Kaplan DI, Buettner SW, Li D, Huang S, Koster van Groos PG, Jaffé PR, et al. In situ porewater uranium concentrations in a contaminated wetland: Effect of seasons and sediment depth. Appl Geochemistry [Internet]. 2017 Oct 1 [cited 2019 Jan 27];85:128–36. Available from: https://www.sciencedirect.com/science/article/pii/S0883292716303626#tbl1
10. O'Brien SL, Gibbons SM, Owens SM, Hampton-Marcell J, Johnston ER, Jastrow JD, et al. Spatial scale drives patterns in soil bacterial diversity. Environ Microbiol. 2016;18(6):2039–51.
11. High-Speed, Multiplexed 16S Microbial Sequencing on the MiSeq® System [Internet]. 2011 [cited 2019 Oct 4]. Available from: http://qiime.org/tutorials/processing_illumina_data.html
12. Feng G, Xie T, Wang X, Bai J, Tang L, Zhao H, et al. Metagenomic analysis of microbial community and function involved in cd-contaminated soil. BMC Microbiol. 2018 Feb 13;18(1).
13. Tipayno SC, Truu J, Sandipan Samaddar |, Truu M, Preem J-K, Kristjan Oopkaup |, et al. The bacterial community structure and functional profile in the heavy metal contaminated paddy soils, surrounding a nonferrous smelter in South Korea. Ecol Evol [Internet].

2018;8:6157. Available from: www.ecolevol.org

14. Li X, Meng D, Li J, Yin H, Liu H, Liu X, et al. Response of soil microbial communities and microbial interactions to long-term heavy metal contamination. Environ Pollut. 2017;231:908–17.

15. Betancourt A, Looney BB, Savannah River M, Ann Thomas DOE S, River Site S. Tin Distribution and Fate in Tims Branch at the Savannah River Site [Internet]. 2011 [cited 2019 Jun 11]. Available from: https://fellows.fiu.edu/wp-content/uploads/2015/06/Amaury-Betancourt-Internship-Rpt-Summer-2011-Rev-0-1.pdf

16. Qiagen. DNeasy PowerSoil Kit Handbook. Qiagen [Internet]. 2016;(June):1–1. Available from: https://www.qiagen.com/de/resources/resourcedetail?id=5a0517a7-711d-4085-8a28-2bb25fab828a&lang=en

17. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012 Aug;6(8):1621–4.

18. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011 Mar 15;108(SUPPL. 1):4516–22.

19. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;

20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Vol. 7, Nature Methods. 2010. p. 335–6.

21. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience [Internet]. 2013 Nov 26 [cited 2019 Oct 23];2(1):16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24280061

# APPENDIX A.

**Table 9. Sequencing mapping file used for downstream analysis**

| #SampleID | BarcodeSequence | LinkerPrimerSequence | Plate | Treatment | Soil_type | Description |
|---|---|---|---|---|---|---|
| JM-CTRL-1 | GTCCGCAAGTTA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A1 | Control_1 | Sand | JM-CTRL-1 |
| JM-CTRL-2 | CAACACATGCTG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A2 | Control_2 | Sand | JM-CTRL-2 |
| JM-CTRL-3 | CATACCGTGAGT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A3 | Control_3 | Sand | JM-CTRL-3 |
| JM-CTRL-4 | GTCCATGGTTCG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A4 | Control_4 | Sand | JM-CTRL-4 |
| JM-CTRL-5 | ACCATTACCATT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A5 | Control_5 | Sand | JM-CTRL-5 |
| JM-CTRL-6 | TGGTAAGAGTCT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A6 | Control_6 | Sand | JM-CTRL-6 |
| JM-CTRL-7 | CCAGCCTTCAGA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A7 | Control_7 | Sand | JM-CTRL-7 |
| JM-CTRL-8 | ATTCAGATGGCA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A8 | Control_8 | Sand | JM-CTRL-8 |
| JM-RR-1 | TTATTCTCTAGG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A9 | Low_1 | Loam | JM-RR-1 |
| JM-RR-2 | TTCGTGAGGATA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A10 | Low_2 | Loam | JM-RR-2 |
| JM-RR-3 | GCGTCATGCATC | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A11 | Low_3 | Loam | JM-RR-3 |
| JM-RR-4 | CCTCGGGTACTA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_A12 | Low_4 | Loam | JM-RR-4 |
| JM-RR-5 | CTACTAGCGGTA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B1 | Low_5 | Loam | JM-RR-5 |
| JM-RR-6 | CGATTTAGGCCA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B2 | Low_6 | Loam | JM-RR-6 |
| JM-RR-7 | GCTTGGTAGGTT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B3 | Low_7 | Loam | JM-RR-7 |
| JM-RR-8 | AGGCGCTCTCCT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B4 | Low_8 | Loam | JM-RR-8 |
| JM-UPS-1 | ACCTGATCCGCA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B5 | Mid_1 | Loam | JM-UPS-1 |
| JM-UPS-2 | GAGATTTAAGCA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B6 | Mid_2 | Loam | JM-UPS-2 |
| JM-UPS-3 | TGGGTCCCACAT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B7 | Mid_3 | Loam | JM-UPS-3 |
| JM-UPS-4 | ATTCTGCCGAAG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B8 | Mid_4 | Loam | JM-UPS-4 |
| JM-UPS-5 | TTGCCTGGGTCA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B9 | Mid_5 | Loam | JM-UPS-5 |
| JM-UPS-6 | TCGTAAGCCGTC | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B10 | Mid_6 | Loam | JM-UPS-6 |
| JM-UPS-7 | ATTAGATTGGAG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B11 | Mid_7 | Loam | JM-UPS-7 |
| JM-UPS-8 | TTAGCCCAGCGT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_B12 | Mid_8 | Loam | JM-UPS-8 |
| JM-LWS-1 | ACTAGGATCAGT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C1 | High_1 | Loamy_sand | JM-LWS-1 |
| JM-LWS-2 | TACACCTTACCT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C2 | High_2 | Loamy_sand | JM-LWS-2 |
| JM-LWS-3 | AGTGTCGATTCG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C3 | High_3 | Loamy_sand | JM-LWS-3 |
| JM-LWS-4 | ATCTCGCTGGGT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C4 | High_4 | Loamy_sand | JM-LWS-4 |
| JM-LWS-5 | ATTCCATTTAGA | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C5 | High_5 | Loamy_sand | JM-LWS-5 |
| JM-LWS-6 | CTGCTGGGAAGG | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C6 | High_6 | Loamy_sand | JM-LWS-6 |
| JM-LWS-7 | TCCTCTTTGGTC | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C7 | High_7 | Loamy_sand | JM-LWS-7 |
| JM-LWS-8 | CAGACTTTCATT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C8 | High_8 | Loamy_sand | JM-LWS-8 |
| BLK-1 | CTTTGGGCCGCT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C9 | Blank_1 | na | BLK-1 |
| BLK-2 | TCCTCGAGCGAT | GTGTGYCAGCMGCCGCGGTAA | Primer_Plate15_C10 | Blank_2 | na | BLK-2 |

**Table 10. Summary of the sequences using split libraries to demultiplex the reads for all samples.**

```
Input file paths
Mapping filepath: Mapping_file.txt (md5: 037ff91e484a6b1606f0296febed6a29)
Sequence read filepath: fastqjoin.join.fastq (md5: 25b547d09113c235b518ec28bda258a6)
Barcode read filepath: fastqjoin.join_barcodes.fastq (md5: e97557871760cfe3f4702184a1a38035)

Quality filter results
Total number of input sequences: 14512909
Barcode not in mapping file: 11784812
Read too short after quality truncation: 37098
Count of N characters exceeds limit: 0
Illumina quality digit = 0: 0
Barcode errors exceed max: 0

Result summary (after quality filtering)
Median sequence length: 253.00
JM-LWS-8        123632
JM-LWS-7        110447
JM-RR-7 108950
JM-RR-1 107892
JM-UPS-6        102189
JM-LWS-1        101190
JM-UPS-2        98353
JM-RR-6 97494
JM-CTRL-6       96030
JM-UPS-7        94281
JM-RR-4 93768
JM-UPS-3        92817
JM-CTRL-2       91961
JM-CTRL-8       88194
JM-CTRL-7       86459
JM-LWS-6        84589
JM-CTRL-4       84398
JM-LWS-3        84381
JM-LWS-2        79937
JM-UPS-1        79420
JM-UPS-5        78042
JM-LWS-4        77972
JM-CTRL-3       76720
JM-RR-8 76467
JM-RR-2 75388
JM-RR-5 71052
JM-LWS-5        69251
JM-CTRL-1       66969
JM-CTRL-5       61369
JM-UPS-4        50258
JM-UPS-8        43665
JM-RR-3 23213
BLK-2   13531
BLK-1   720

Total number seqs written       2690999
```

w ---