Phylogenetic Analysis of Bacteria in Water Samples Using the QIIME2 Metagenomics Platform

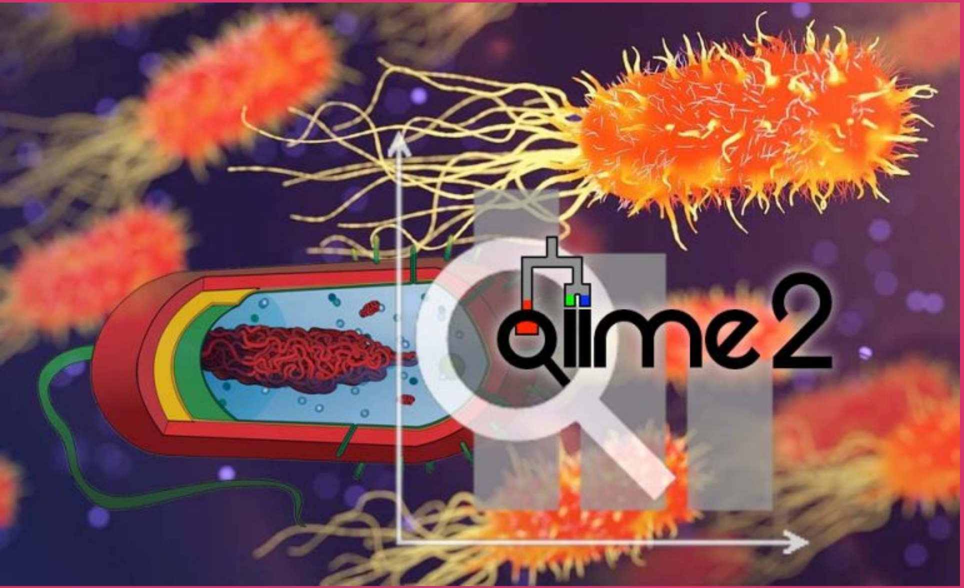
---

Elliot Dean **Jalen Bailey** **Denver Baptiste, PhD**

Stevens Institute of Technology

2 June 2023

What is QIIME2?



https://qiime2.org

https://doi.org/10.1038/s41587-019-0209-9

- A free, open source bioinformatics platform
- Focused on microbiome analysis
- Community developed, with dozens of standard tools and extra plugins

What Does QIIME Mean?

Quantitative

Insights

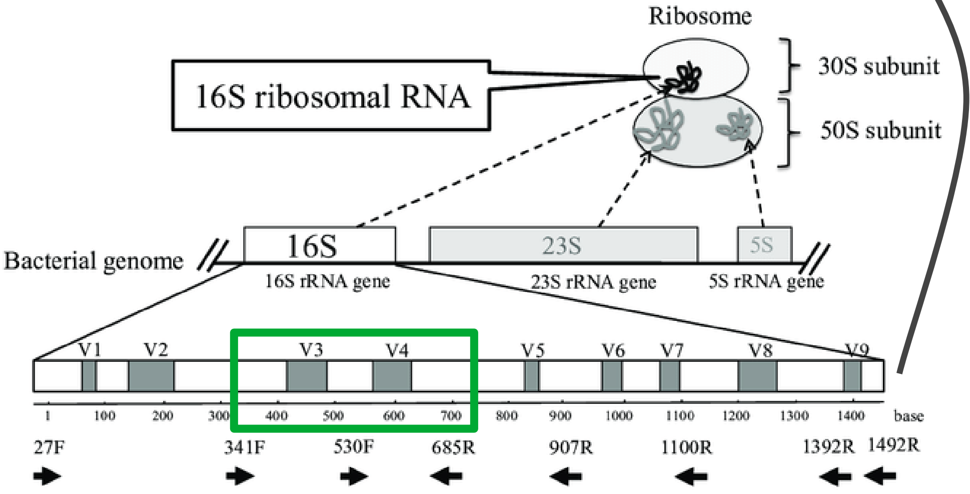
Into

Microbial

Ecology

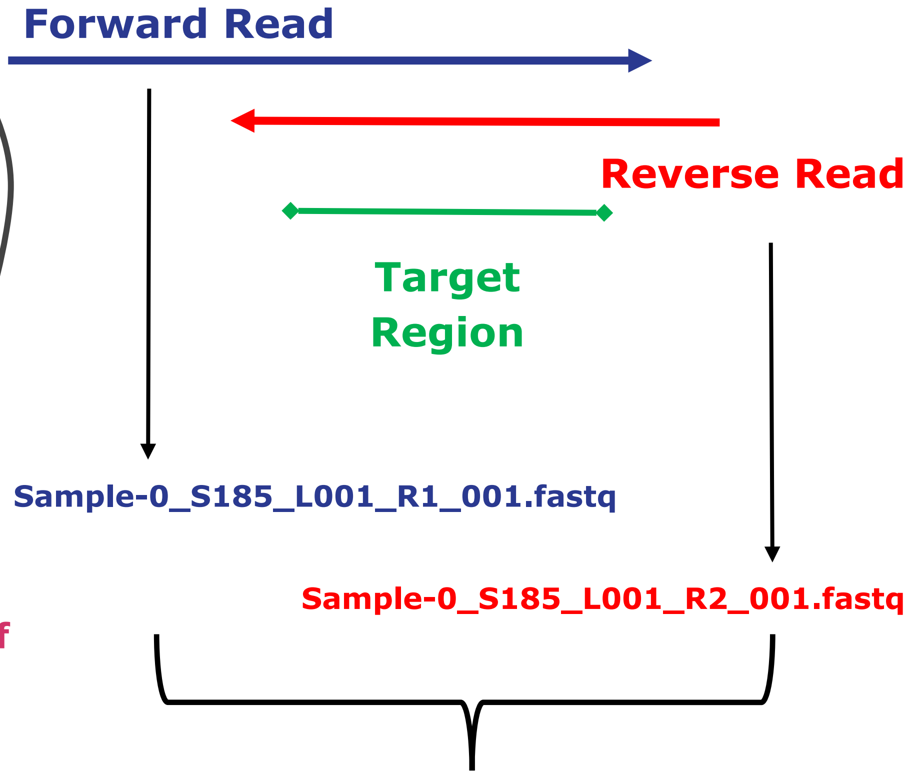
2 = Version 2

16S Ribosomal RNA Sequencing



- **16S rRNA is a component of the 30S subunit of the ribosome complex**
- **The majority of the gene is conserved across species**
- **However, 9 variable regions exist**

V Region = ? bp



QIIME2 combines the reads and processes their data to return the sequence of the target region

Workflow Overview

I. **Samples of wastewater are collected from location of interest**

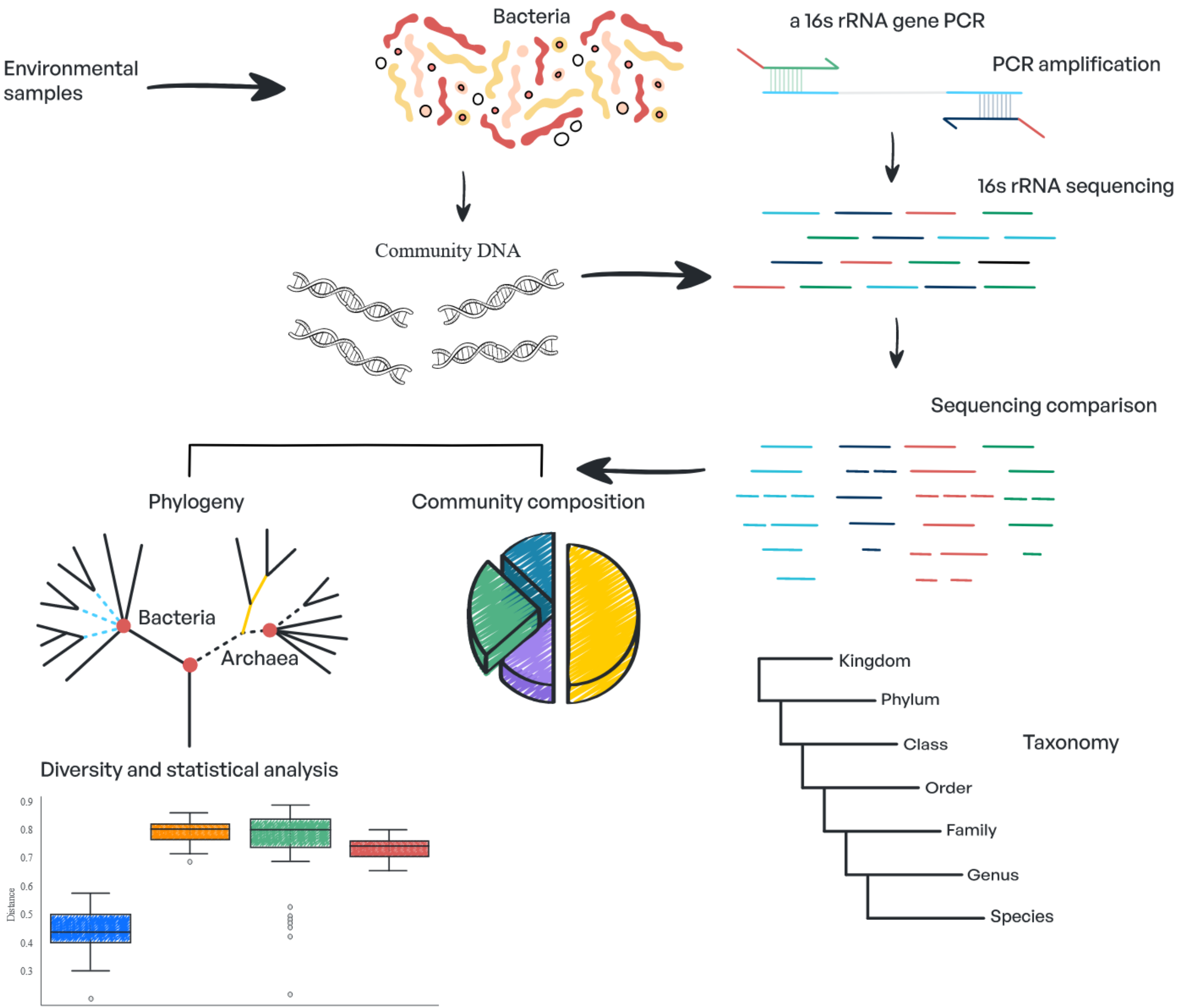
II. **Bacteria and other organisms are harvested**

III. **Genomic DNA (containing the 16S gene) is purified**

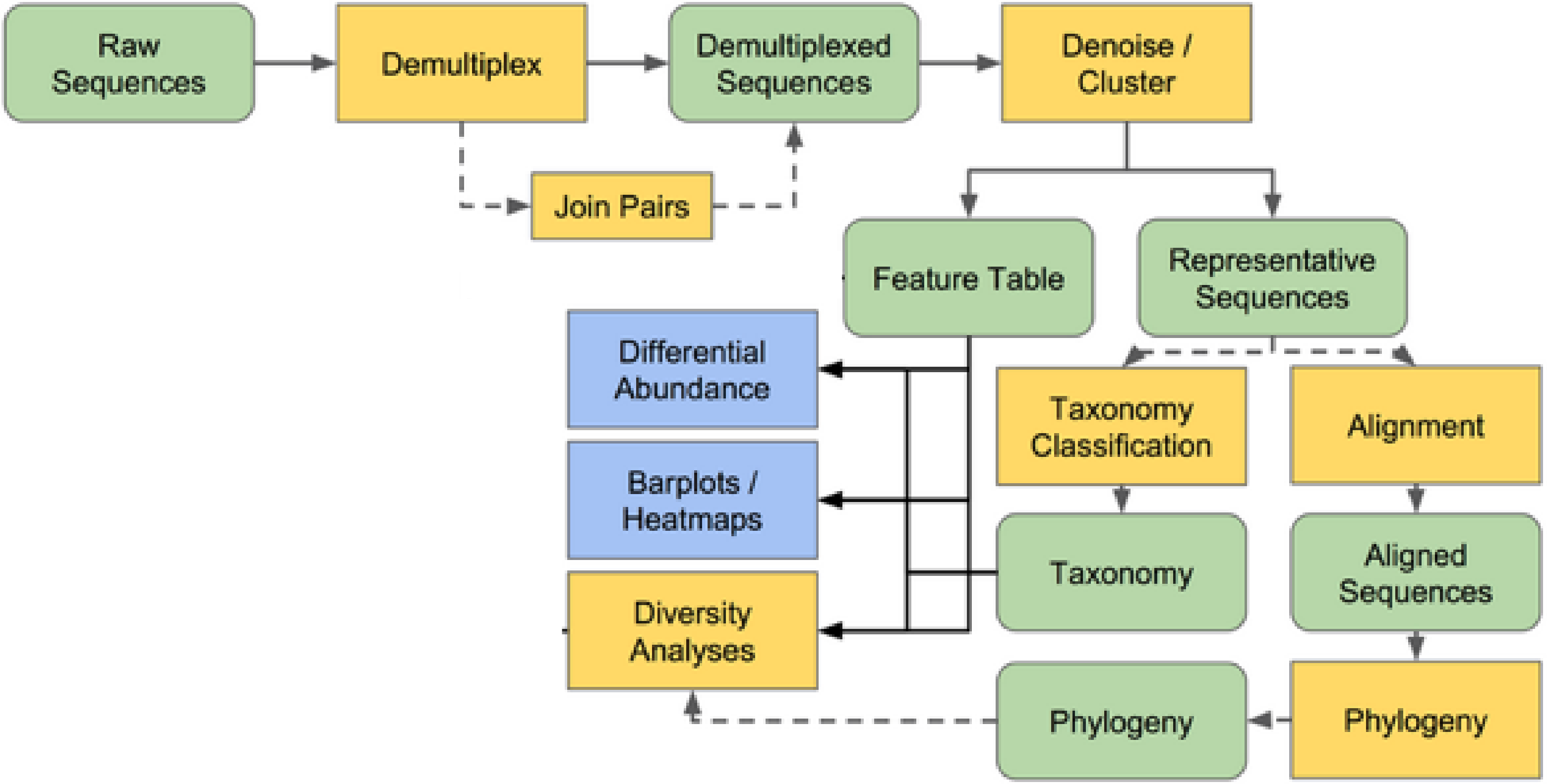
IV. **Primers targeting variable regions, including barcode sequences, are used to amplify DNA via PCR**

V. **Amplicons are sent for sequencing**

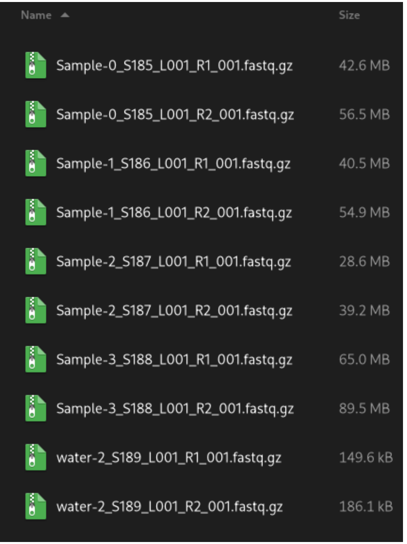
VI. **Data is reviewed, and metagenomic analysis proceeds**



Qiime2 Pipeline

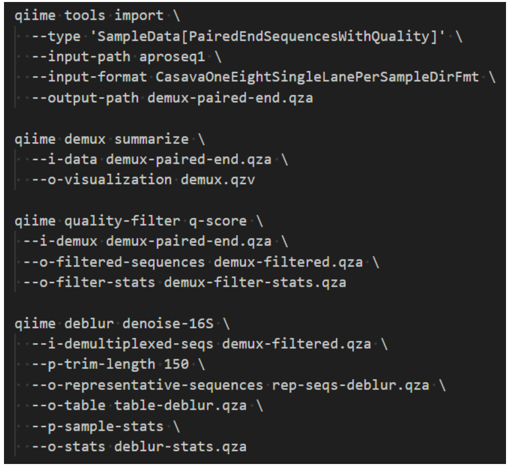


Qiime2 Datatypes



Name ▲	Size
Sample-0_S185_L001_R1_001.fastq.gz	42.6 MB
Sample-0_S185_L001_R2_001.fastq.gz	56.5 MB
Sample-1_S186_L001_R1_001.fastq.gz	40.5 MB
Sample-1_S186_L001_R2_001.fastq.gz	54.9 MB
Sample-2_S187_L001_R1_001.fastq.gz	28.6 MB
Sample-2_S187_L001_R2_001.fastq.gz	39.2 MB
Sample-3_S188_L001_R1_001.fastq.gz	65.0 MB
Sample-3_S188_L001_R2_001.fastq.gz	89.5 MB
water-2_S189_L001_R1_001.fastq.gz	149.6 kB
water-2_S189_L001_R2_001.fastq.gz	186.1 kB

Sequence Files from Illumina



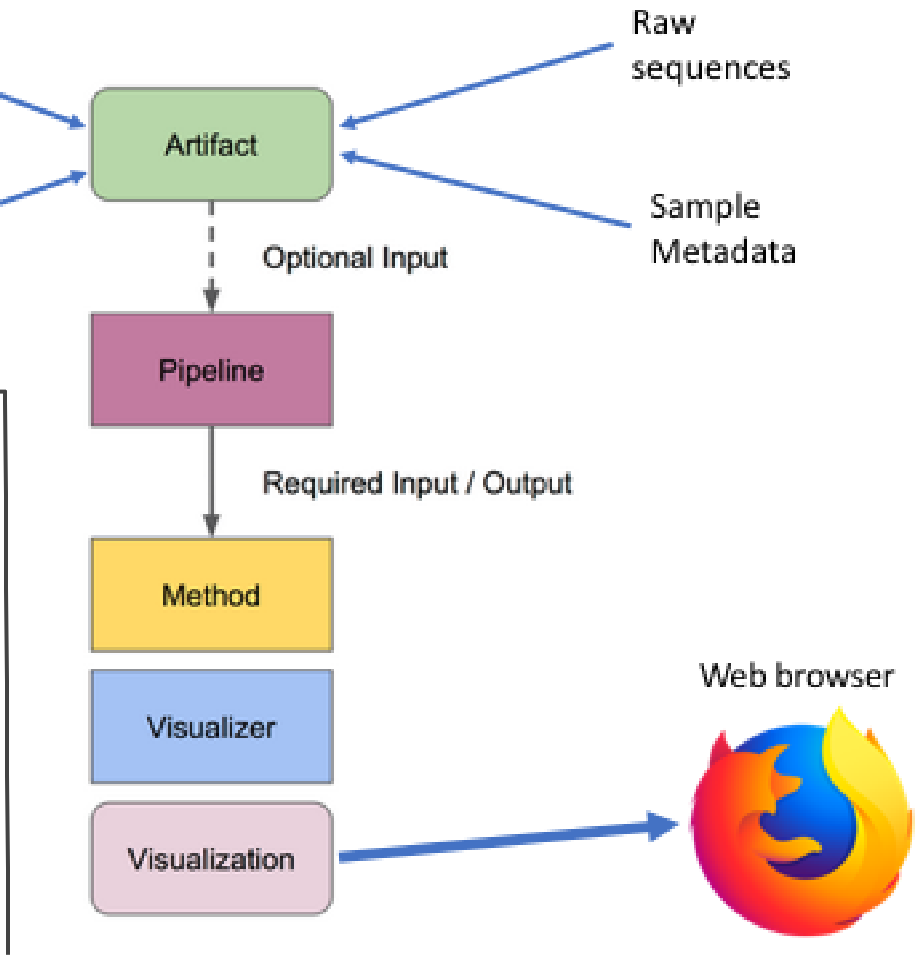
```
qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path aproseq1 \
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \
  --output-path demux-paired-end.qza

qiime demux summarize \
  --i-data demux-paired-end.qza \
  --o-visualization demux.qzv

qiime quality-filter q-score \
 --i-demux demux-paired-end.qza \
 --o-filtered-sequences demux-filtered.qza \
 --o-filter-stats demux-filter-stats.qza

qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 150 \
  --o-representative-sequences rep-seqs-deblur.qza \
  --o-table table-deblur.qza \
  --p-sample-stats \
  --o-stats deblur-stats.qza
```

Example Methods



Reference sequences

Reference taxonomy

Raw sequences

Sample Metadata

Artifact

Optional Input

Pipeline

Required Input / Output

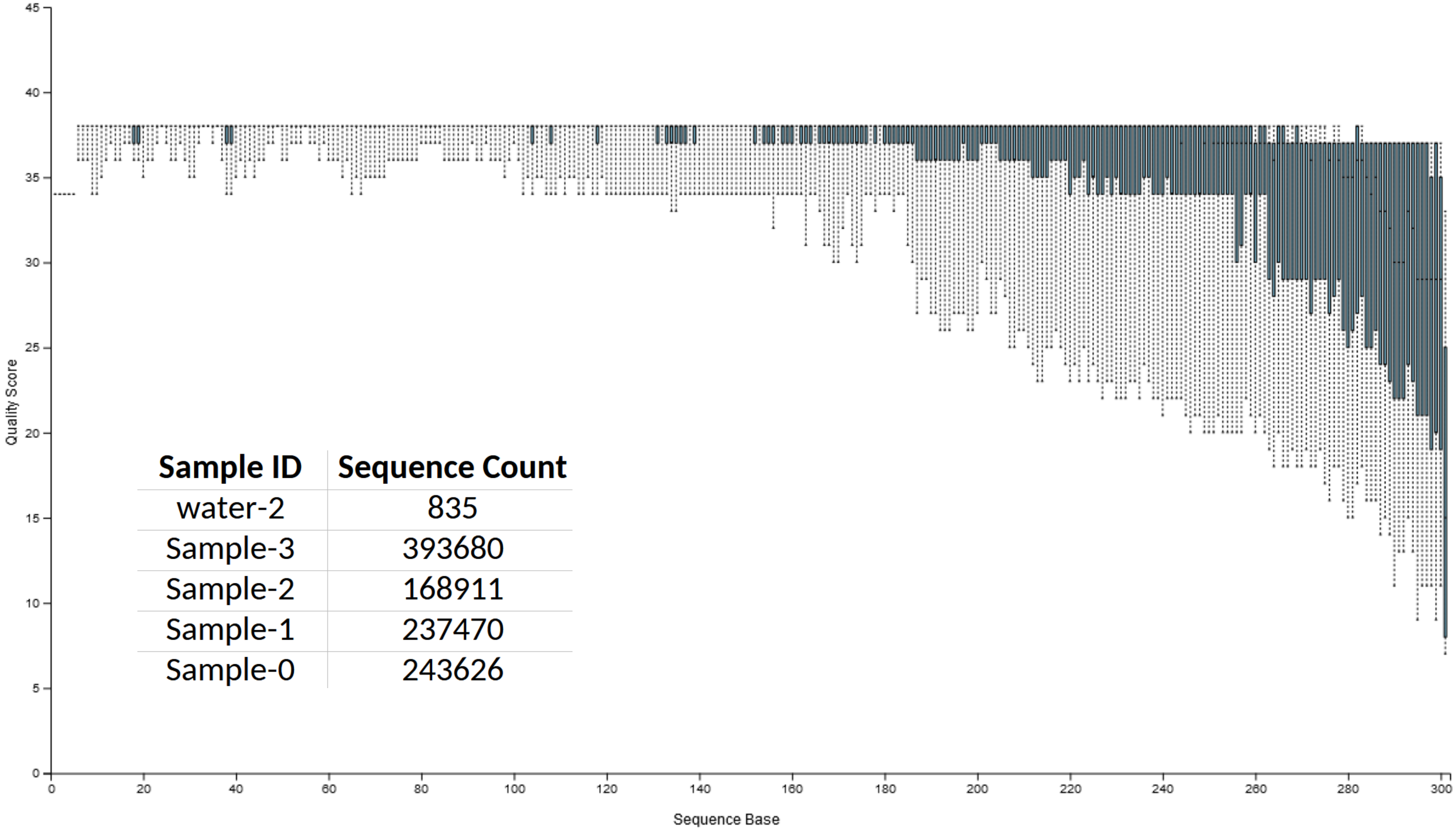
Method

Visualizer

Visualization

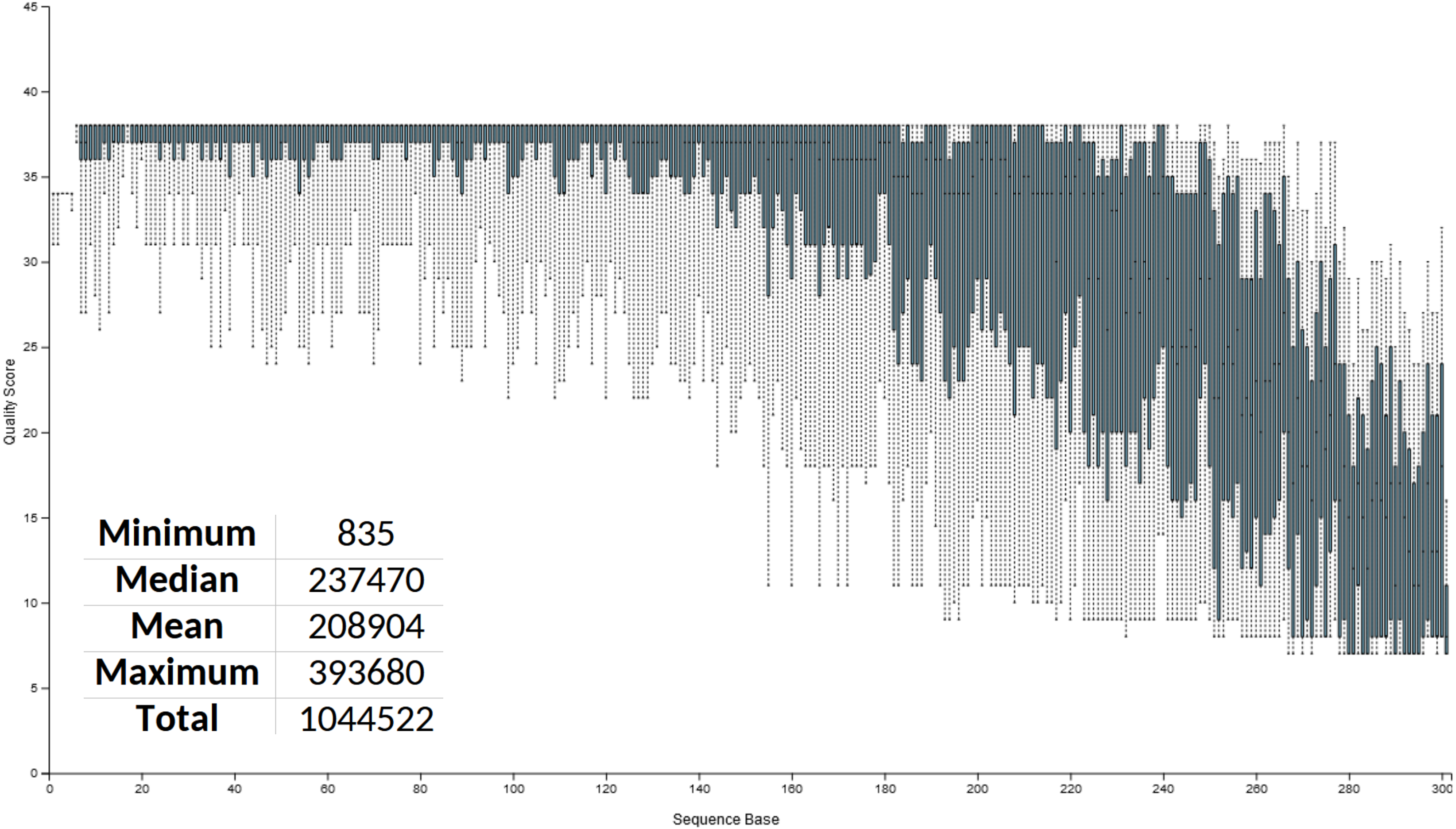
Web browser

Demultiplexed Forward Read Quality Plot



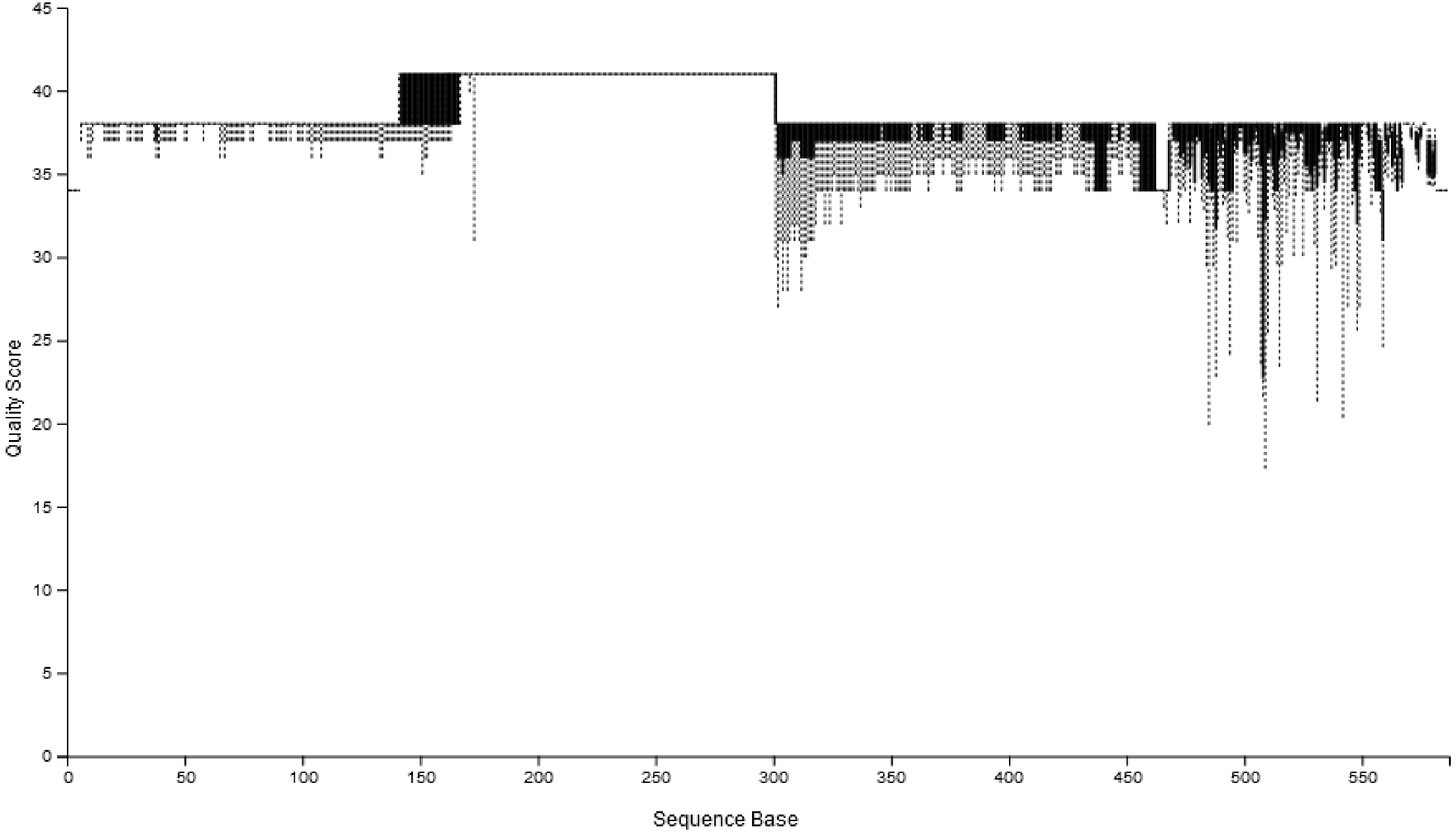
Sample ID	Sequence Count
water-2	835
Sample-3	393680
Sample-2	168911
Sample-1	237470
Sample-0	243626

Demultiplexed Reverse Read Quality Plot

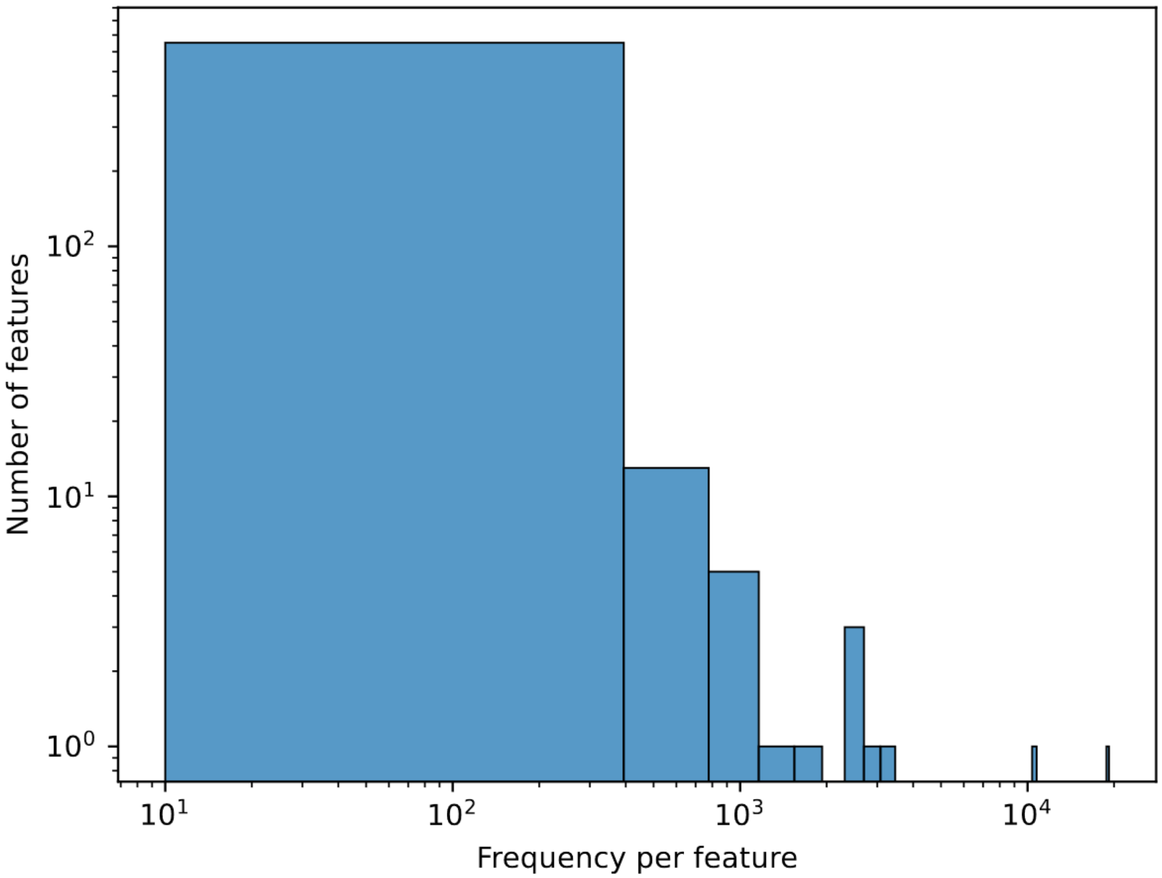


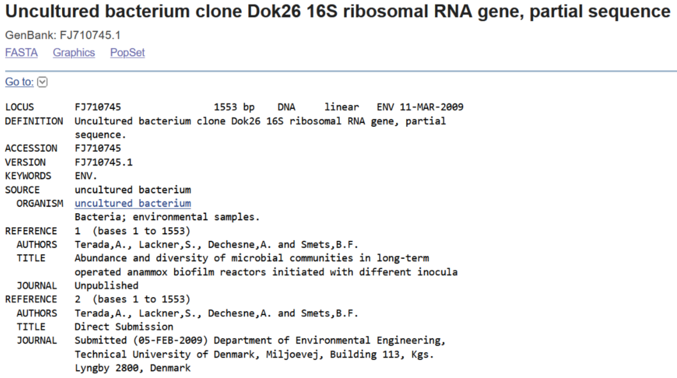
Minimum	835
Median	237470
Mean	208904
Maximum	393680
Total	1044522

Joined Demultiplexed Reads Quality Plot



Features Identified in Sample Sequences

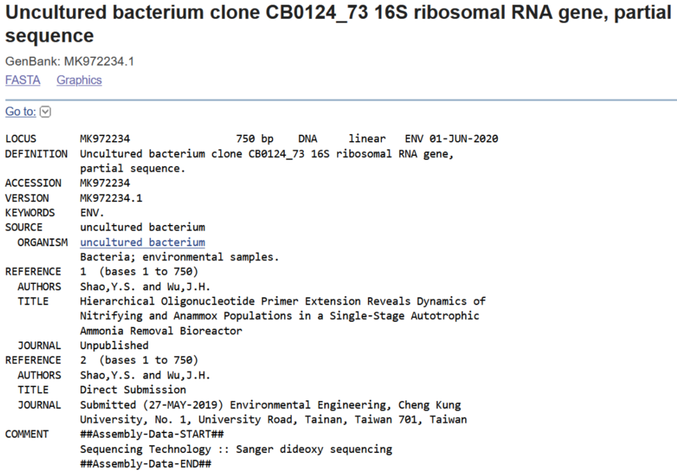




Uncultured bacterium clone Dok26 16S ribosomal RNA gene, partial sequence

GenBank: FJ710745.1

```
LOCUS       FJ710745                1553 bp    DNA     linear   ENV 11-MAR-2009
DEFINITION  Uncultured bacterium clone Dok26 16S ribosomal RNA gene, partial
            sequence.
ACCESSION   FJ710745
VERSION     FJ710745.1
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 1553)
  AUTHORS   Terada,A., Lackner,S., Dechesne,A. and Smets,B.F.
  TITLE     Abundance and diversity of microbial communities in long-term
            operated anammox biofilm reactors initiated with different inocula
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1553)
  AUTHORS   Terada,A., Lackner,S., Dechesne,A. and Smets,B.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-FEB-2009) Department of Environmental Engineering,
            Technical University of Denmark, Miljoevej, Building 113, Kgs.
            Lyngby 2800, Denmark
```

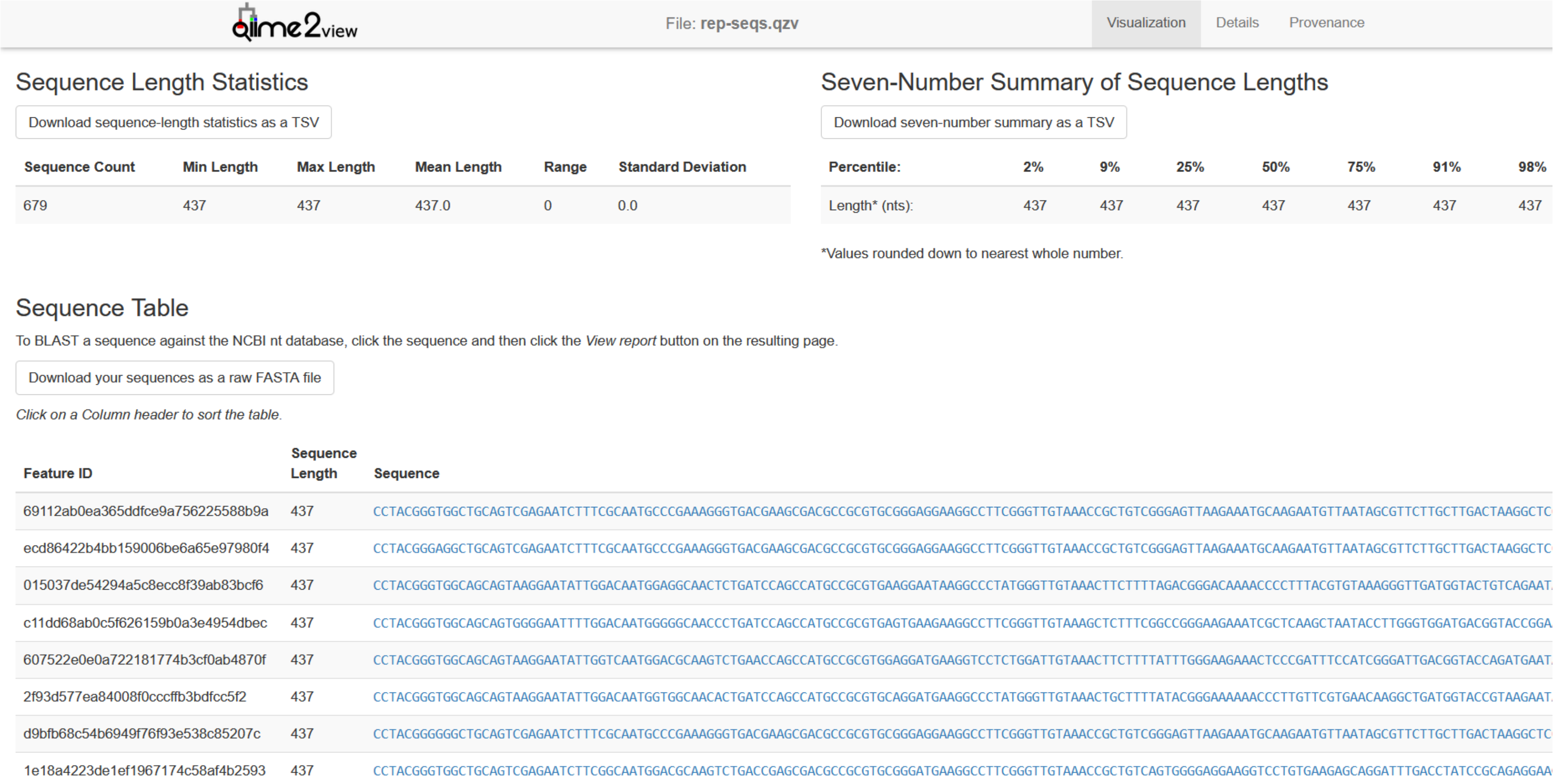


Uncultured bacterium clone CB0124_73 16S ribosomal RNA gene, partial sequence

GenBank: MK972234.1

```
LOCUS       MK972234                 750 bp    DNA     linear   ENV 01-JUN-2020
DEFINITION  Uncultured bacterium clone CB0124_73 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   MK972234
VERSION     MK972234.1
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 750)
  AUTHORS   Shao,Y.S. and Wu,J.H.
  TITLE     Hierarchical Oligonucleotide Primer Extension Reveals Dynamics of
            Nitrifying and Anammox Populations in a Single-Stage Autotrophic
            Ammonia Removal Bioreactor
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 750)
  AUTHORS   Shao,Y.S. and Wu,J.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-MAY-2019) Environmental Engineering, Cheng Kung
            University, No. 1, University Road, Tainan, Taiwan 701, Taiwan
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
```

Identified Feature Sequences



File: **rep-seqs.qzv**

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
679	437	437	437.0	0	0.0

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	437	437	437	437	437	437	437

*Values rounded down to nearest whole number.

Sequence Table

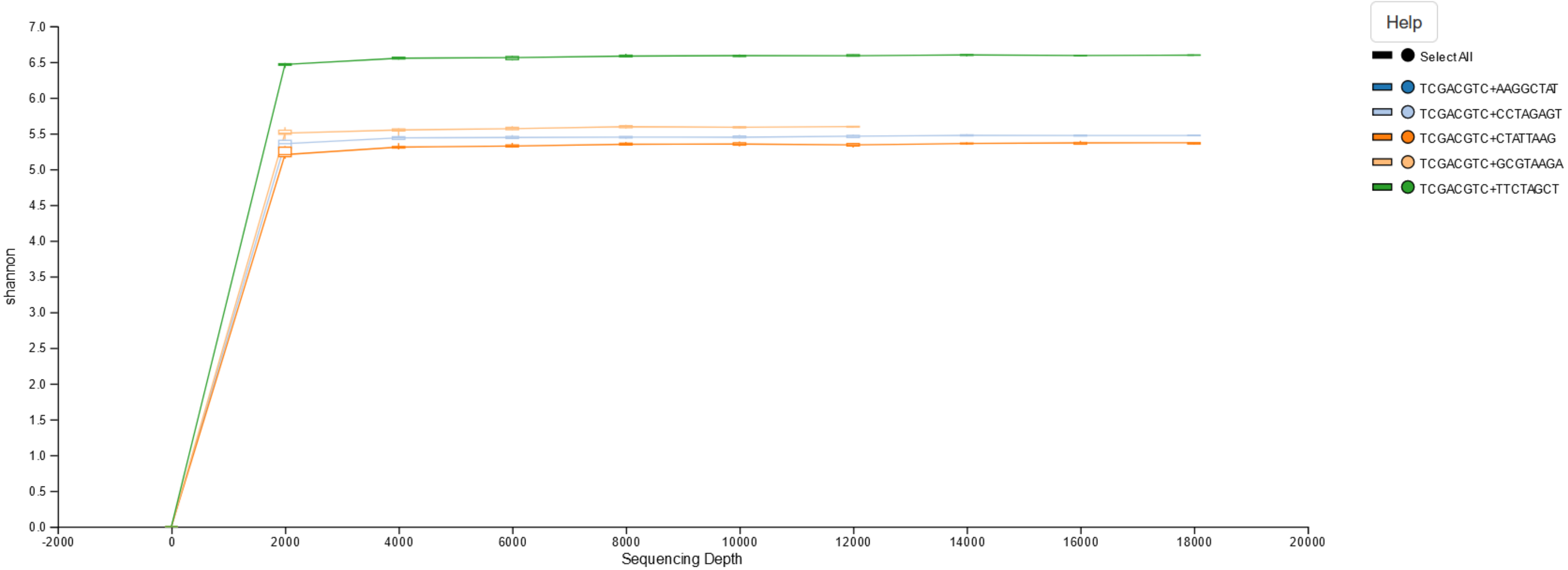
To BLAST a sequence against the NCBI nt database, click the sequence and then click the *View report* button on the resulting page.

Download your sequences as a raw FASTA file

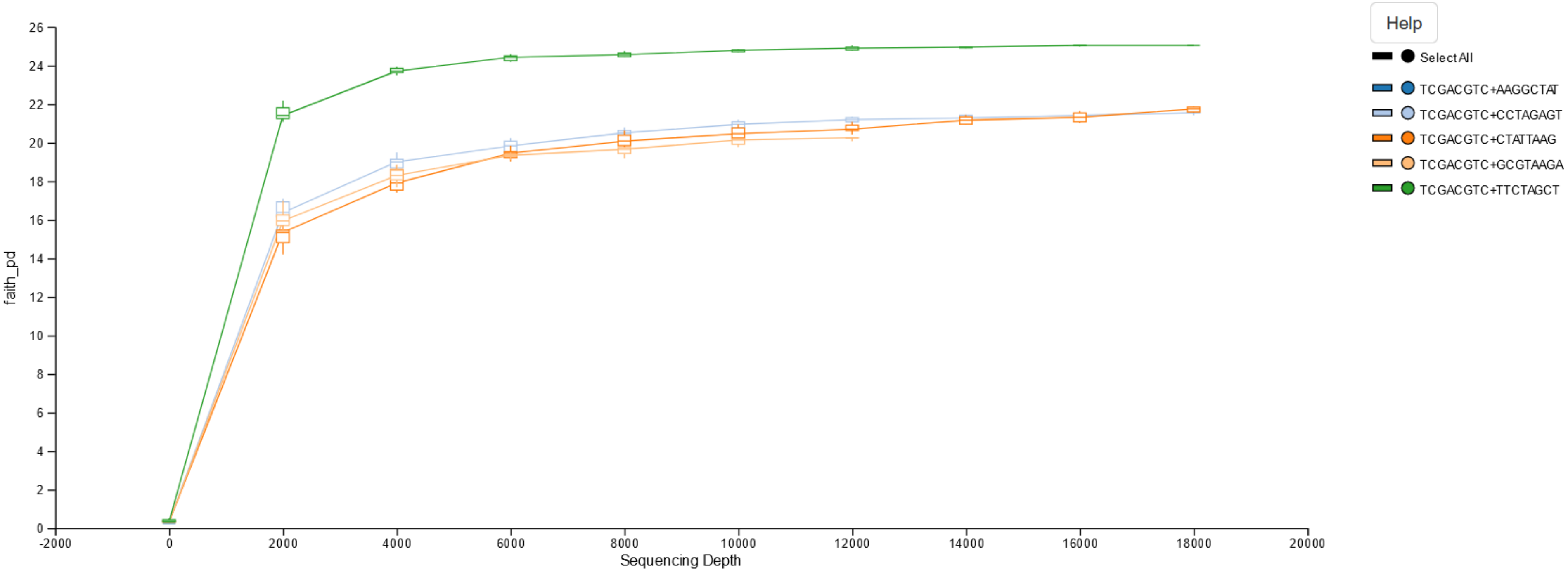
Click on a Column header to sort the table.

Feature ID	Sequence Length	Sequence
69112ab0ea365ddfce9a756225588b9a	437	CCTACGGGTGGCTGCAGTCGAGAATCTTTCGCAATGCCCGAAAGGGTGACGAAGCGACGCCGCGTGCGGGAGGAAGGCCTTCGGGTTGTAAACCGCTGTCGGGAGTTAAGAAATGCAAGAATGTTAATAGCGTTCTTGCTTGACTAAGGCTC
ecd86422b4bb159006be6a65e97980f4	437	CCTACGGGAGGCTGCAGTCGAGAATCTTTCGCAATGCCCGAAAGGGTGACGAAGCGACGCCGCGTGCGGGAGGAAGGCCTTCGGGTTGTAAACCGCTGTCGGGAGTTAAGAAATGCAAGAATGTTAATAGCGTTCTTGCTTGACTAAGGCTC
015037de54294a5c8ecc8f39ab83bcf6	437	CCTACGGGTGGCAGCAGTAAGGAATATTGGACAATGGAGGCAACTCTGATCCAGCCATGCCGCGTGAAGGAATAAGGCCCTATGGGTTGTAAACTTCTTTTAGACGGGACAAAACCCCTTTACGTGTAAAGGGTTGATGGTACTGTCAGAAT
c11dd68ab0c5f626159b0a3e4954dbec	437	CCTACGGGTGGCAGCAGTGGGGAATTTTGGACAATGGGGGCAACCCTGATCCAGCCATGCCGCGTGAGTGAAGAAGGCCTTCGGGTTGTAAAGCTCTTTCGGCCGGGAAGAAATCGCTCAAGCTAATACCTTGGGTGGATGACGGTACCGGA
607522e0e0a722181774b3cf0ab4870f	437	CCTACGGGTGGCAGCAGTAAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGGAGGATGAAGGTCCTCTGGATTGTAAACTTCTTTTATTTGGGAAGAAACTCCCGATTTCCATCGGGATTGACGGTACCAGATGAAT
2f93d577ea84008f0cccffb3bdfcc5f2	437	CCTACGGGTGGCAGCAGTAAGGAATATTGGACAATGGTGGCAACACTGATCCAGCCATGCCGCGTGCAGGATGAAGGCCCTATGGGTTGTAAACTGCTTTTATACGGGAAAAAACCCTTGTTCGTGAACAAGGCTGATGGTACCGTAAGAAT
d9bfb68c54b6949f76f93e538c85207c	437	CCTACGGGGGGCTGCAGTCGAGAATCTTTCGCAATGCCCGAAAGGGTGACGAAGCGACGCCGCGTGCGGGAGGAAGGCCTTCGGGTTGTAAACCGCTGTCGGGAGTTAAGAAATGCAAGAATGTTAATAGCGTTCTTGCTTGACTAAGGCTC
1e18a4223de1ef1967174c58af4b2593	437	CCTACGGGTGGCTGCAGTCGAGAATCTTCGGCAATGGACGCAAGTCTGACCGAGCGACGCCGCGTGCGGGATGAAGGCCTTCGGGTTGTAAACCGCTGTCAGTGGGGAGGAAGGTCCTGTGAAGAGCAGGATTTGACCTATCCGCAGAGGAA

Measuring the Diversity of Species Within a Community

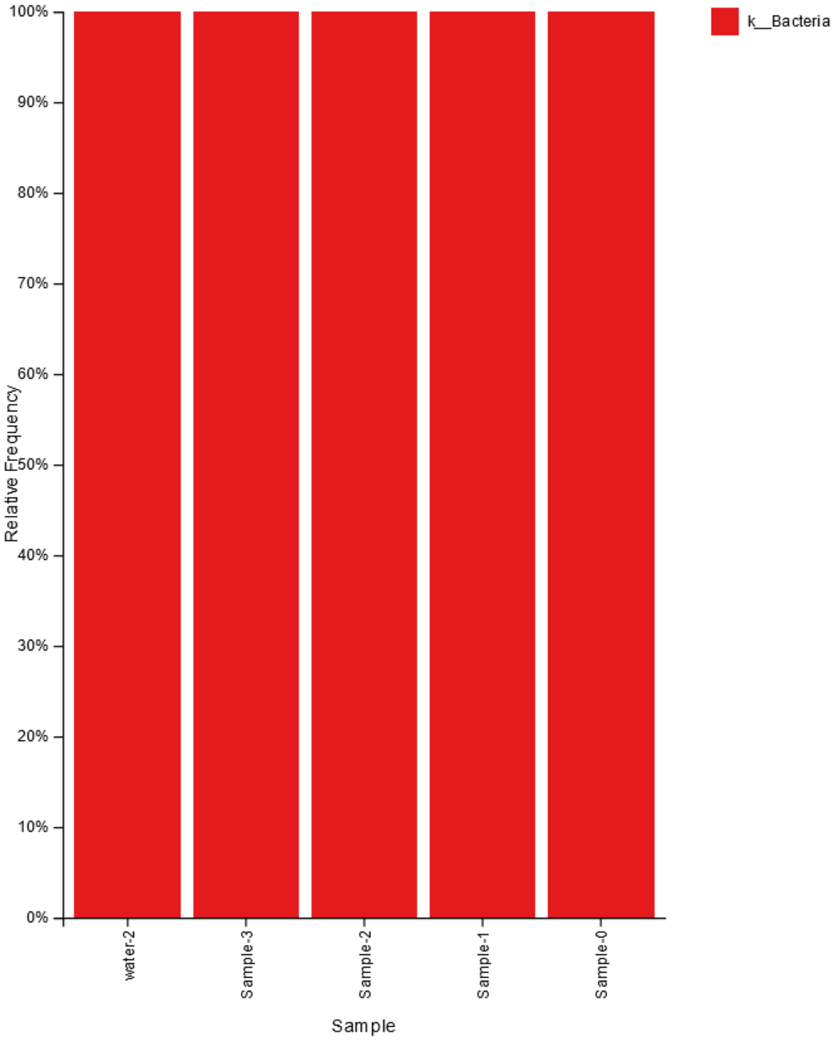


Branch Length: Connecting all Species in Samples



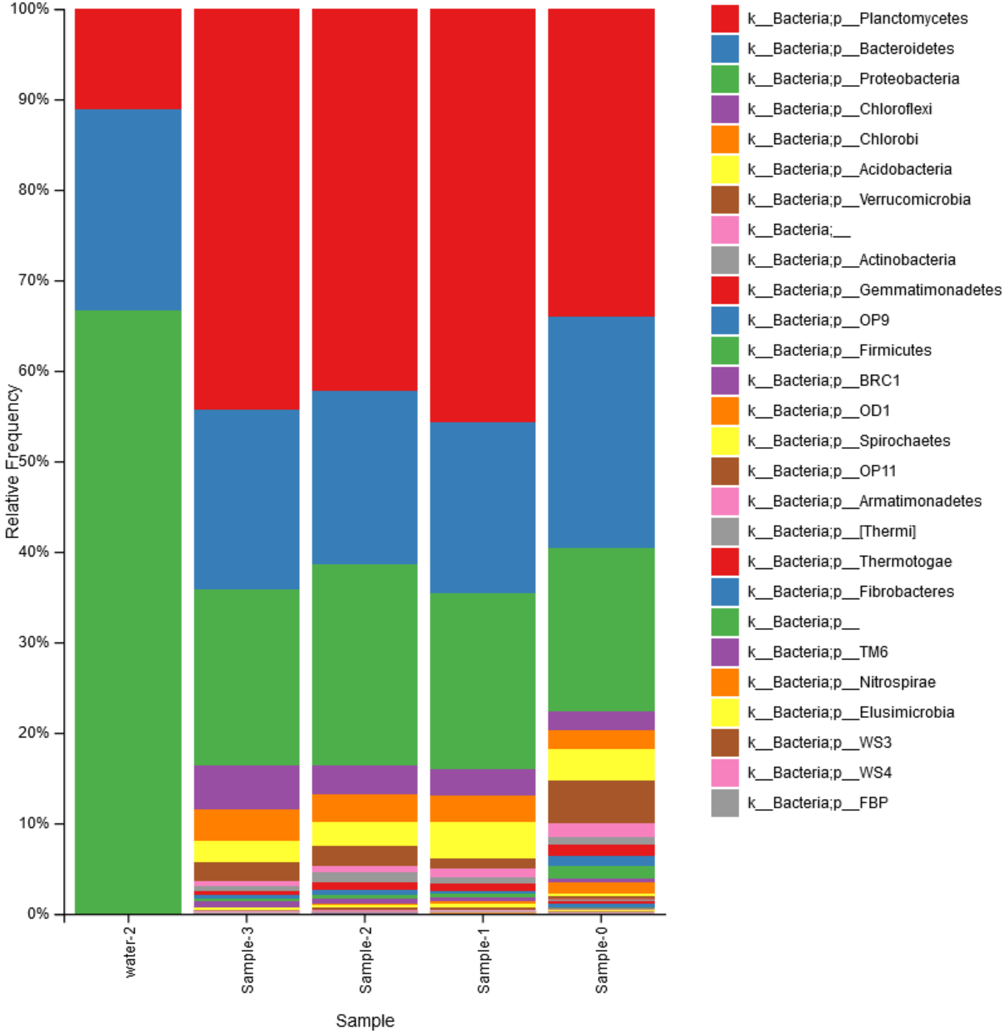
Taxonomy: Level 1

- Each level of a phylogenetic tree has a certain number of branches
- At level 1, the only distinction between samples is Domain
- As such, every bar is nearly identical
- The more branches that are traversed, the more specific the taxa become

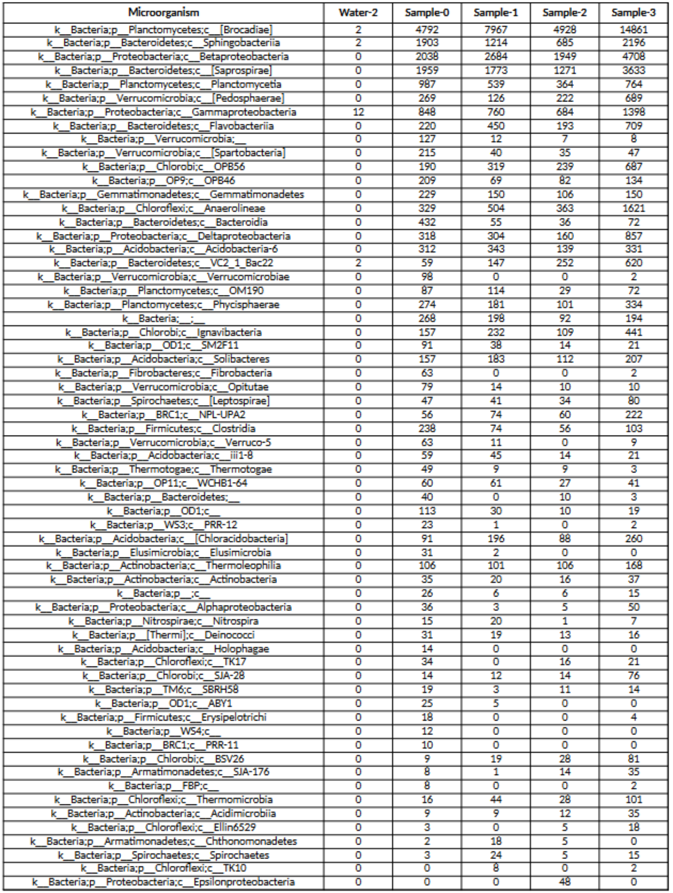


Taxonomy: Level 2

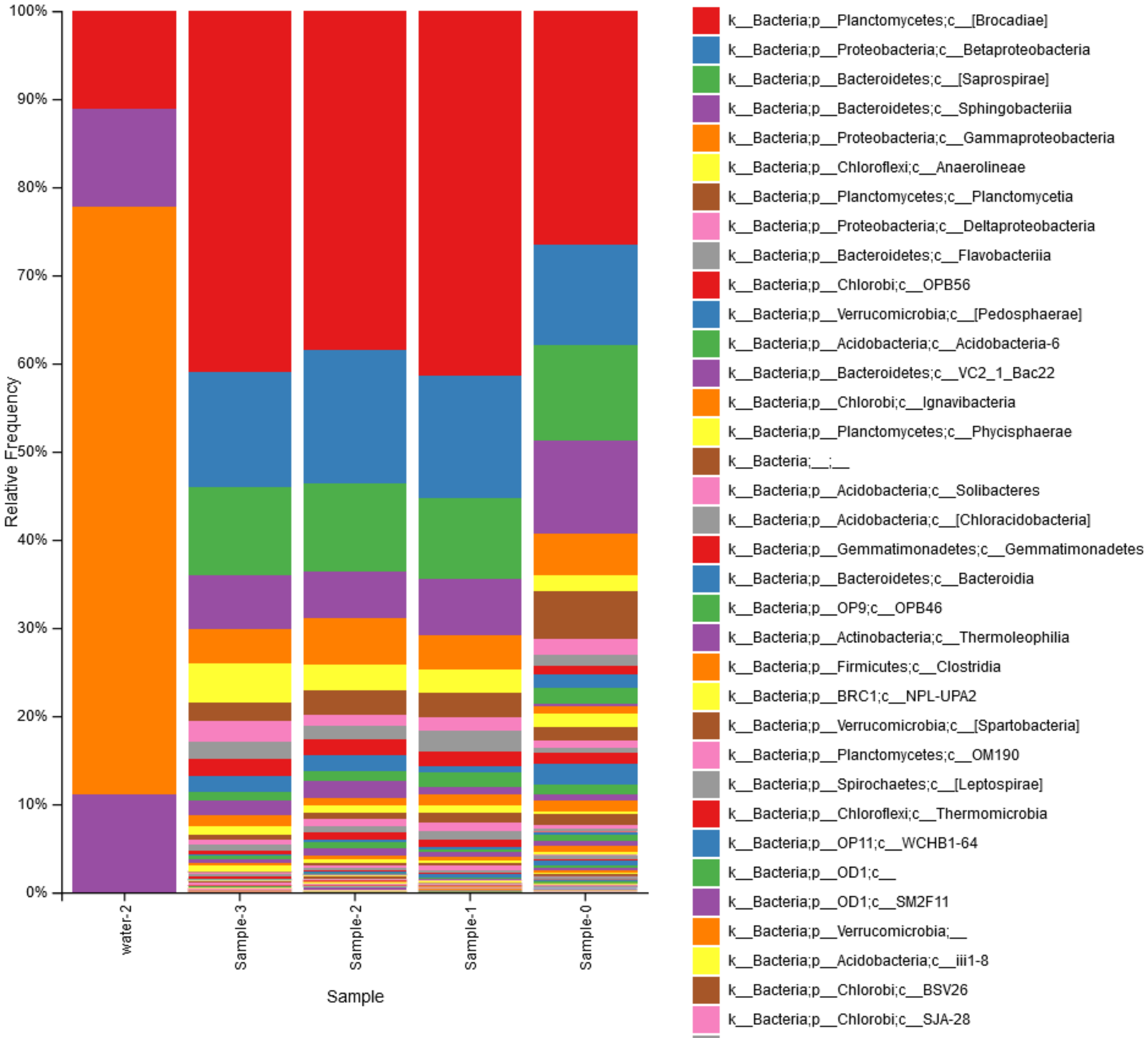
Microorganism	Water-2	Sample-0	Sample-1	Sample-2	Sample-3
k__Bacteria;p__Planctomycetes	2	6140	8801	5422	16031
k__Bacteria;p__Bacteroidetes	4	4613	3639	2447	7233
k__Bacteria;p__Proteobacteria	12	3240	3751	2846	7013
k__Bacteria;p__Verrucomicrobia	0	851	203	274	765
k__Bacteria;p__Chlorobi	0	370	582	390	1285
k__Bacteria;p__OP9	0	209	69	82	134
k__Bacteria;p__Gemmatimonadetes	0	229	150	106	150
k__Bacteria;p__Chloroflexi	0	382	556	412	1763
k__Bacteria;p__Acidobacteria	0	633	767	353	819
k__Bacteria;__	0	268	198	92	194
k__Bacteria;p__OD1	0	229	73	24	40
k__Bacteria;p__Fibrobacteres	0	63	0	0	2
k__Bacteria;p__Spirochaetes	0	50	65	39	95
k__Bacteria;p__BRC1	0	66	74	60	222
k__Bacteria;p__Firmicutes	0	256	74	56	107
k__Bacteria;p__Thermotogae	0	49	9	9	3
k__Bacteria;p__OP11	0	60	61	27	41
k__Bacteria;p__WS3	0	23	1	0	2
k__Bacteria;p__Elusimicrobia	0	31	2	0	0
k__Bacteria;p__Actinobacteria	0	150	130	134	240
k__Bacteria;p__	0	26	6	6	15
k__Bacteria;p__Nitrospirae	0	15	20	1	7
k__Bacteria;p__[Thermi]	0	31	19	13	16
k__Bacteria;p__TM6	0	19	3	11	14
k__Bacteria;p__WS4	0	12	0	0	0
k__Bacteria;p__Armatimonadetes	0	10	19	19	35
k__Bacteria;p__FBP	0	8	0	0	2



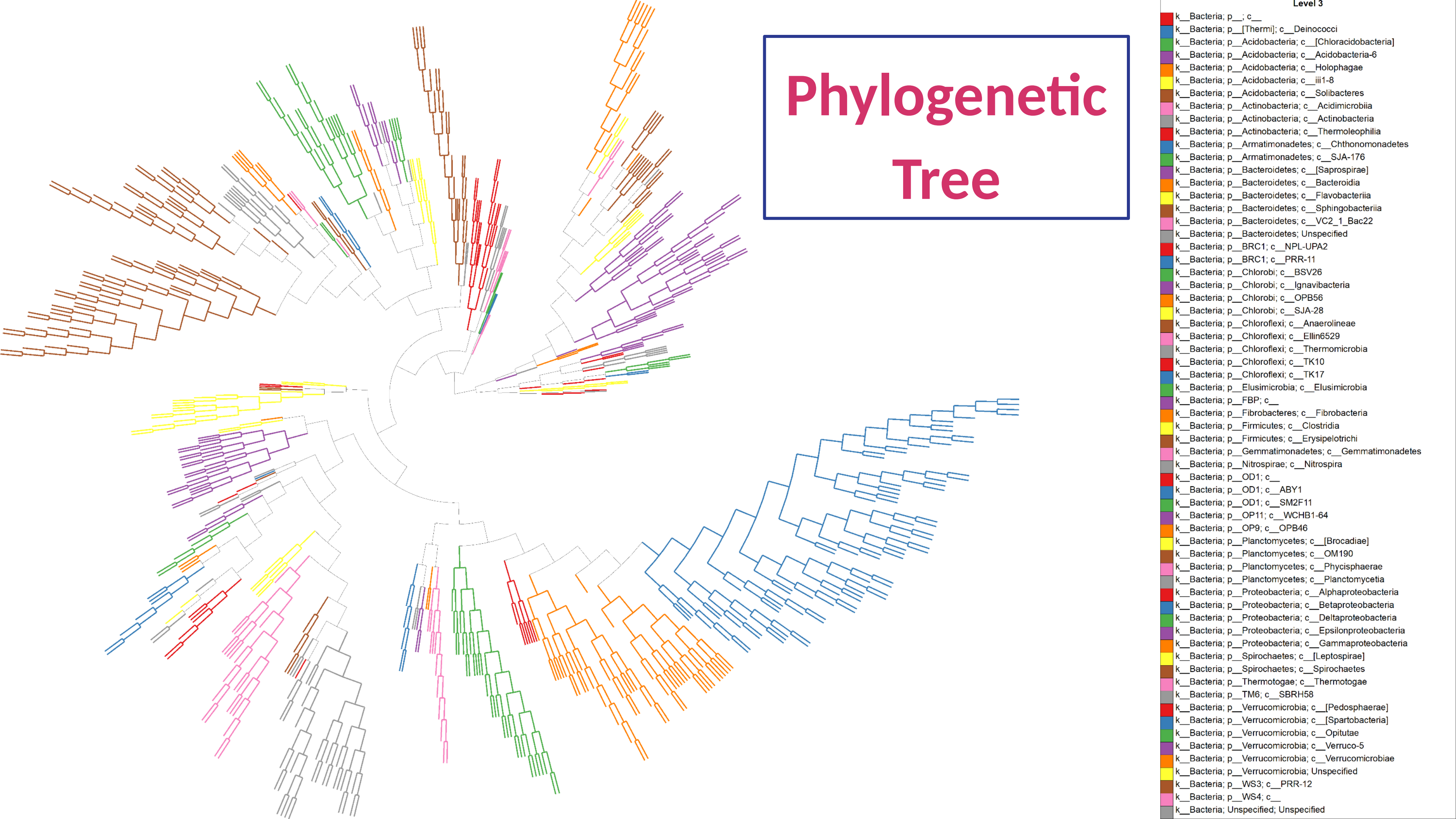
Taxonomy: Level 3



Microorganism	Water-2	Sample-0	Sample-1	Sample-2	Sample-3
k__Bacteria;p__Planctomycetes;c__[Brocadiae]	2	4792	7967	4928	14861
k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia	2	1903	1214	685	2196
k__Bacteria;p__Proteobacteria;c__Betaproteobacteria	0	2038	2684	1949	4708
k__Bacteria;p__Bacteroidetes;c__[Saprospirae]	0	1959	1773	1271	3633
k__Bacteria;p__Planctomycetes;c__Planctomycetia	0	987	539	364	764
k__Bacteria;p__Verrucomicrobia;c__[Pedosphaerae]	0	269	126	222	689
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria	12	848	760	684	1398
k__Bacteria;p__Bacteroidetes;c__Flavobacteriia	0	220	450	193	709
k__Bacteria;p__Verrucomicrobia;__	0	127	12	7	8
k__Bacteria;p__Verrucomicrobia;c__[Spartobacteria]	0	215	40	35	47
k__Bacteria;p__Chlorobi;c__OPB56	0	190	319	239	687
k__Bacteria;p__OP9;c__OPB46	0	209	69	82	134
k__Bacteria;p__Gemmatimonadetes;c__Gemmatimonadetes	0	229	150	106	150
k__Bacteria;p__Chloroflexi;c__Anaerolineae	0	329	504	363	1621
k__Bacteria;p__Bacteroidetes;c__Bacteroidia	0	432	55	36	72
k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria	0	318	304	160	857
k__Bacteria;p__Acidobacteria;c__Acidobacteria-6	0	312	343	139	331
k__Bacteria;p__Bacteroidetes;c__VC2_1_Bac22	2	59	147	252	620
k__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae	0	98	0	0	2
k__Bacteria;p__Planctomycetes;c__OM190	0	87	114	29	72
k__Bacteria;p__Planctomycetes;c__Phycisphaerae	0	274	181	101	334
k__Bacteria;__;__	0	268	198	92	194
k__Bacteria;p__Chlorobi;c__Ignavibacteria	0	157	232	109	441
k__Bacteria;p__OD1;c__SM2F11	0	91	38	14	21
k__Bacteria;p__Acidobacteria;c__Solibacteres	0	157	183	112	207
k__Bacteria;p__Fibrobacteres;c__Fibrobacteria	0	63	0	0	2
k__Bacteria;p__Verrucomicrobia;c__Opitutae	0	79	14	10	10
k__Bacteria;p__Spirochaetes;c__[Leptospirae]	0	47	41	34	80
k__Bacteria;p__BRC1;c__NPL-UPA2	0	56	74	60	222
k__Bacteria;p__Firmicutes;c__Clostridia	0	238	74	56	103
k__Bacteria;p__Verrucomicrobia;c__Verruco-5	0	63	11	0	9
k__Bacteria;p__Acidobacteria;c__iii1-8	0	59	45	14	21
k__Bacteria;p__Thermotogae;c__Thermotogae	0	49	9	9	3
k__Bacteria;p__OP11;c__WCHB1-64	0	60	61	27	41
k__Bacteria;p__Bacteroidetes;__	0	40	0	10	3
k__Bacteria;p__OD1;c__	0	113	30	10	19
k__Bacteria;p__WS3;c__PRR-12	0	23	1	0	2
k__Bacteria;p__Acidobacteria;c__[Chloracidobacteria]	0	91	196	88	260
k__Bacteria;p__Elusimicrobia;c__Elusimicrobia	0	31	2	0	0
k__Bacteria;p__Actinobacteria;c__Thermoleophilia	0	106	101	106	168
k__Bacteria;p__Actinobacteria;c__Actinobacteria	0	35	20	16	37
k__Bacteria;p__;c__	0	26	6	6	15
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria	0	36	3	5	50
k__Bacteria;p__Nitrospirae;c__Nitrospira	0	15	20	1	7
k__Bacteria;p__[Thermi];c__Deinococci	0	31	19	13	16
k__Bacteria;p__Acidobacteria;c__Holophagae	0	14	0	0	0
k__Bacteria;p__Chloroflexi;c__TK17	0	34	0	16	21
k__Bacteria;p__Chlorobi;c__SJA-28	0	14	12	14	76
k__Bacteria;p__TM6;c__SBRH58	0	19	3	11	14
k__Bacteria;p__OD1;c__ABY1	0	25	5	0	0
k__Bacteria;p__Firmicutes;c__Erysipelotrichi	0	18	0	0	4
k__Bacteria;p__WS4;c__	0	12	0	0	0
k__Bacteria;p__BRC1;c__PRR-11	0	10	0	0	0
k__Bacteria;p__Chlorobi;c__BSV26	0	9	19	28	81
k__Bacteria;p__Armatimonadetes;c__SJA-176	0	8	1	14	35
k__Bacteria;p__FBP;c__	0	8	0	0	2
k__Bacteria;p__Chloroflexi;c__Thermomicrobia	0	16	44	28	101
k__Bacteria;p__Actinobacteria;c__Acidimicrobiia	0	9	9	12	35
k__Bacteria;p__Chloroflexi;c__Ellin6529	0	3	0	5	18
k__Bacteria;p__Armatimonadetes;c__Chthonomonadetes	0	2	18	5	0
k__Bacteria;p__Spirochaetes;c__Spirochaetes	0	3	24	5	15
k__Bacteria;p__Chloroflexi;c__TK10	0	0	8	0	2
k__Bacteria;p__Proteobacteria;c__Epsilonproteobacteria	0	0	0	48	0

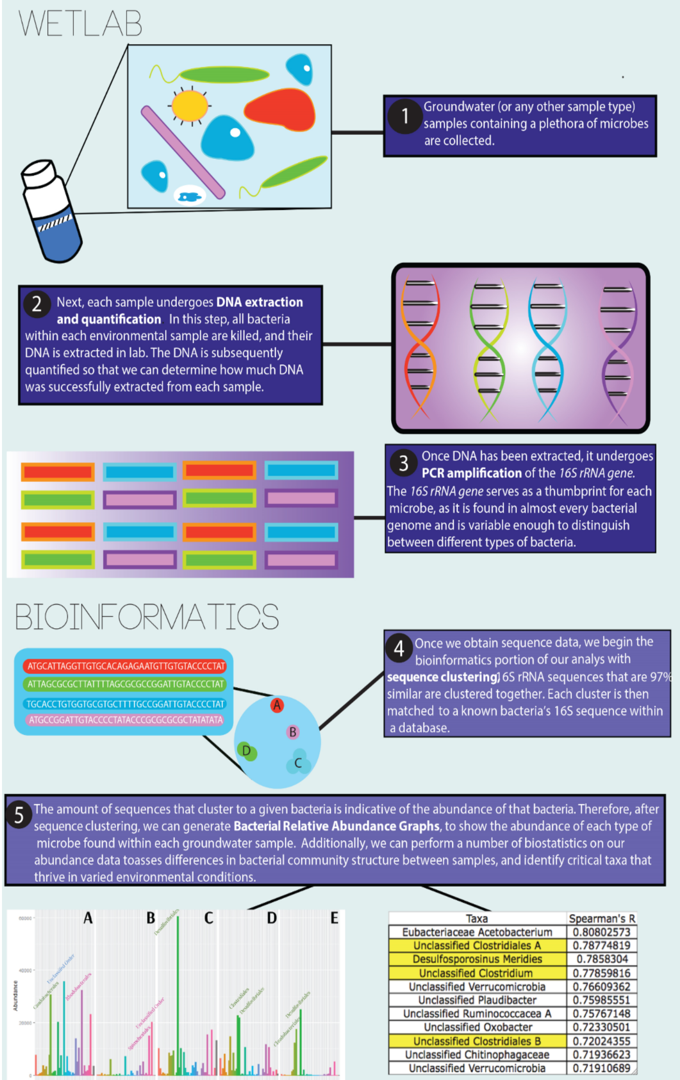


Phylogenetic Tree



Conclusion

- **QIIME2 provides a comprehensive framework for metagenomic analysis**
- **A plethora of tools are easily available to import different data types, perform quality control, filter / trim sequences, and more**
- **Files generated by QIIME2 are easily shared, and can be opened in most web browsers**
- **The QIIME2 documentation contains thorough tutorials and guides, ideal for novices**



WETLAB

1 Groundwater (or any other sample type) samples containing a plethora of microbes are collected.

2 Next, each sample undergoes **DNA extraction and quantification** In this step, all bacteria within each environmental sample are killed, and their DNA is extracted in lab. The DNA is subsequently quantified so that we can determine how much DNA was successfully extracted from each sample.

3 Once DNA has been extracted, it undergoes **PCR amplification** of the *16S rRNA gene*. The *16S rRNA gene* serves as a thumbprint for each microbe, as it is found in almost every bacterial genome and is variable enough to distinguish between different types of bacteria.

BIOINFORMATICS

ATGCATTAGGTTGTGCACAGAGAATGTTGTGTACCCCTAT
ATTAGCGCGCTTATTTTAGCGCGCCGGATTGTACCCCTAT
TGCACCTGTGGTGCGTGCTTTTGCCGGATTGTACCCCTAT
ATGCCGGATTGTACCCCTATACCCGCGCGCGCTATATATA

4 Once we obtain sequence data, we begin the bioinformatics portion of our analys with **sequence clustering**16S rRNA sequences that are 97% similar are clustered together. Each cluster is then matched to a known bacteria's 16S sequence within a database.

5 The amount of sequences that cluster to a given bacteria is indicative of the abundance of that bacteria. Therefore, after sequence clustering, we can generate **Bacterial Relative Abundance Graphs**, to show the abundance of each type of microbe found within each groundwater sample. Additionally, we can perform a number of biostatistics on our abundance data toasses differences in bacterial community structure between samples, and identify critical taxa that thrive in varied environmental conditions.

Taxa	Spearman's R
Eubacteriaceae Acetobacterium	0.80802573
Unclassified Clostridiales A	0.78774819
Desulfosporosinus Meridies	0.7858304
Unclassified Clostridium	0.77859816
Unclassified Verrucomicrobia	0.76609362
Unclassified Plaudibacter	0.75985551
Unclassified Ruminococcacea A	0.75767148
Unclassified Oxobacter	0.72330501
Unclassified Clostridiales B	0.72024355
Unclassified Chitinophagaceae	0.71936623
Unclassified Verrucomicrobia	0.71910689

Thank You For Listening!

Any Questions?