




# Amplicons, Metagenomes, and Metatranscriptomes from Sediment and Water

Madison R. Newman,<sup>a</sup> Darlenys Sanchez,<sup>a</sup> Anna M. Acosta,<sup>a</sup>  Bernadette J. Connors<sup>a</sup>

<sup>a</sup>Science Department, Dominican University New York, Orangeburg, New York, USA

**ABSTRACT** High fecal indicator bacterium (FIB) counts in water have been found to correlate with high sediment FIB counts. To determine the other bacterial populations in common between the two substrates, sediment and water samples from suburban waters known to be impacted by stormwater runoff were examined using next-generation sequencing.

Sites in the lower Hudson River watershed were initially chosen based on data obtained from Hudson Riverkeeper (1), as well as one site that was not included in their analyses (Spring Valley, NY). Riverkeeper is a nonprofit environmental organization dedicated to the protection of the Hudson River and its tributaries. Several sites that were sampled that had high failure rates, as determined by whether the samples collected previously by Riverkeeper met the EPA guideline for safe swimming. Water samples (1 L) were collected in sterile Nalgene bottles that were first rinsed with creek water three times prior to being fully submerged. The water was filtered through sterile nitrocellulose filters (pore size, 0.22  $\mu$ m). Nearshore creek bed sediment (5 mL) was collected by submerging closed, sterile 15-mL conical tubes and then releasing the seal to collect the sediment, making all efforts to minimize water flow into the collection bottle.

DNA and RNA were extracted from 0.25 g of each sample using the ZymoBIOMICS DNA/RNA miniprep kit. Metagenomic libraries were constructed using the Nextera XT DNA library prep kit (Illumina). Metatranscriptomic libraries were prepared with 100 ng of total RNA using the NEBNext Ultra RNA kit for double-stranded cDNA synthesis and metatranscriptome library preparation. Libraries between 250 and 400 bp were purified on a 2% agarose gel using a Qiagen QIAquick gel extraction kit. Sequencing was performed on an Illumina NextSeq 550 instrument at Wright Labs (Huntingdon, PA, USA) to produce 2  $\times$  150-bp reads. FastQC v0.11.9 (2) and fastp v0.22.0 (3) were used to check and filter the raw data. The microbial and functional features of the samples were determined by annotating the paired sequence data using HUMAnN v2 (4), with sequences identified as belonging to *Homo sapiens* removed using KneadData v2 (5). The UNIREF90 (UniProt/UniRef database v2014\_07) genes from the functional annotation were mapped to KEGG v56 orthologs (6). Identification of bacteria to the species level was conducted by collating the HUMAnN v2 taxonomic identifications. Default parameters were used for all software unless otherwise specified.

For 16S rRNA gene microbial community profiling, PCR was performed on DNA extracts based on the Earth Microbiome Project's 16S rRNA gene amplification protocol (7). The PCR products were pooled and purified after separation on a 2% agarose gel. The pooled libraries were quality checked using a 2100 Bioanalyzer high-sensitivity DNA analysis kit (Agilent Technologies). Sequencing was conducted by Wright Labs using Illumina MiSeq v2 chemistry with paired-end 250-bp reads. Demultiplexing was performed using BCL2fastq v2.19.0.316 (Illumina) with default settings. The demultiplexed paired-end reads were processed using QIIME2 v2021.2

**Editor** J. Cameron Thrash, University of Southern California

**Copyright** © 2023 Newman et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Bernadette J. Connors, [bernadette.connors@duney.edu](mailto:bernadette.connors@duney.edu). The authors declare no conflict of interest.

**Received** 21 November 2022

**Accepted** 16 February 2023

**TABLE 1** Properties of the 'omics data sets

Site	GPS coordinates	Substrate	NGS <sup>a</sup> type	No. of reads	SRA accession no.	Predominant taxa (% relative abundance) <sup>b</sup>
Sparkill	41.025363, –73.927466	Sedminet	16S	13,381	<a href="#">SRR22221596</a>	<i>Comamonadaceae</i> (3.4), <i>Dechloromonas</i> (3.4)
			MG	223,012	<a href="#">SRR22221592</a>	Unclassified (100)
			MT	6,832,234	<a href="#">SRR22221579</a>	Unclassified (100)
Sparkill	41.025363, –73.927466	Water	16S	100,660	<a href="#">SRR22221595</a>	<i>Comamonadaceae</i> (39.2), <i>Polynucleobacter</i> (8.9), <i>Dechloromonas</i> (1.1)
			MG	7,240,270	<a href="#">SRR22221591</a>	<i>Polynucleobacter</i> (7.14), unclassified (91.8)
			MT	11,090,123	<a href="#">SRR22221578</a>	<i>Polynucleobacter</i> (4.23), unclassified (95.8)
Blauvelt Arm	41.056438, –73.944968	Sedminet	16S	73,024	<a href="#">SRR22221584</a>	<i>Comamonadaceae</i> (5.9), <i>Dechloromonas</i> (2.8)
			MG	928,109	<a href="#">SRR22221590</a>	Unclassified (100)
			MT	12,648,884	<a href="#">SRR22221577</a>	Unclassified (100)
Blauvelt Arm	41.056438, –73.944968	Water	16S	64,698	<a href="#">SRR22221573</a>	<i>Comamonadaceae</i> (6.8), <i>Polynucleobacter</i> (0.2), <i>Dechloromonas</i> (1.5)
			MG	2,549,195	<a href="#">SRR22221589</a>	<i>Enterobacter</i> (9.5), unclassified (90.5)
			MT	11,661,858	<a href="#">SRR22221576</a>	Unclassified (100)
Marsh	41.038606, –73.915210	Sedminet	16S	61,323	<a href="#">SRR22221566</a>	<i>Comamonadaceae</i> (0.9)
			MG	3,507,848	<a href="#">SRR22221588</a>	<i>Sulfuricella</i> (15.5), unclassified (84.5)
			MT	17,329,967	<a href="#">SRR22221575</a>	Unclassified (100)
Marsh	41.038606, –73.915210	Water	16S	76,475	<a href="#">SRR22221565</a>	<i>Comamonadaceae</i> (5.5), <i>Polynucleobacter</i> (0.1)
			MG	4,019,881	<a href="#">SRR22221587</a>	<i>Flavobacteria</i> (4), <i>Halothiobacillus</i> (2), unclassified (92.7)
			MT	24,687,649	<a href="#">SRR22221574</a>	Unclassified (100)
Moturis	41.015904, –73.937346	Sedminet	16S	80,759	<a href="#">SRR22221562</a>	<i>Comamonadaceae</i> (6.5), <i>Dechloromonas</i> (1.8)
			MG	3,582,410	<a href="#">SRR22221583</a>	<i>Thiobacillus</i> (2.1), unclassified (97.9)
			MT	10,130,592	<a href="#">SRR22221570</a>	Unclassified (100)
Moturis	41.015904, –73.937346	Water	16S	72,182	<a href="#">SRR22221561</a>	<i>Comamonadaceae</i> (11.3), <i>Polynucleobacter</i> (3.5), <i>Dechloromonas</i> (0.5)
			MG	5,739,788	<a href="#">SRR22221582</a>	<i>Enterobacter</i> (4.3), <i>Eubacterium</i> (3.7), <i>Acinetobacter</i> (2.2), <i>Klebsiella</i> (5.3), <i>Polynucleobacter</i> (5.8), <i>Ruminococcus</i> (4.2), unclassified (70.6)
			MT	14,379,033	<a href="#">SRR22221569</a>	<i>Polynucleobacter</i> (4.9), unclassified (95.1)
Spring Valley	41.115367, –74.042263	Sedminet	16S	64,866	<a href="#">SRR22221564</a>	<i>Comamonadaceae</i> (5.2), <i>Dechloromonas</i> (4.3)
			MG	842,136	<a href="#">SRR22221586</a>	Unclassified (100)
			MT	9,706,868	<a href="#">SRR22221572</a>	Unclassified (100)
Spring Valley	41.115367, –74.042263	Water	16S	75,542	<a href="#">SRR22221563</a>	<i>Comamonadaceae</i> (16.7), <i>Polynucleobacter</i> (8.1), <i>Dechloromonas</i> (1.1)
			MG	3,185,050	<a href="#">SRR22221585</a>	<i>Polynucleobacter</i> (8.9), <i>Megamonas</i> (2.7), <i>Microcystis</i> (1.04), unclassified (95.4)
			MT	14,873,372	<a href="#">SRR22221571</a>	<i>Polynucleobacter</i> (4.4), unclassified (86.6)
Rockleigh	41.007620, –73.940000	Sedminet	16S	45,611	<a href="#">SRR22221594</a>	<i>Comamonadaceae</i> (6.1), <i>Dechloromonas</i> (4)
			MG	13,797,720	<a href="#">SRR22221581</a>	<i>Rhodopseudomonas</i> (4.2), <i>Sulfuricella</i> (2.2), <i>Thiobacillus</i> (2.2), unclassified (91.4)
			MT	9,838,983	<a href="#">SRR22221568</a>	<i>Thiobacillus</i> (9.9), unclassified (90.1)
Rockleigh	41.007620, –73.940000	Water	16S	80,467	<a href="#">SRR22221593</a>	<i>Comamonadaceae</i> (35.5), <i>Polynucleobacter</i> (4.3), <i>Dechloromonas</i> (1.4)
			MG	4,825,003	<a href="#">SRR22221580</a>	<i>Polynucleobacter</i> (7.3), unclassified (92.7)
			MT	10,021,470	<a href="#">SRR22221567</a>	<i>Polynucleobacter</i> (4.4), unclassified (95.5)

<sup>a</sup> NGS, next-generation sequencing; MG, metagenomic; MT, metatranscriptomic.

<sup>b</sup> Only select bacterial taxa are reported in this table.

(8) with the DADA2 plug-in (9). The preformatted Silva SSU nonredundant (NR) 99 full-length rRNA gene sequence reference database was used to assign taxonomy (10, 11).

Table 1 details properties of the three 'omics data sets, including the relative abundance of select bacterial taxa. The taxa presented are those that had a relative abundance of >1% and were differentially represented in the two substrates. Although not shown in Table 1, several members of *Bacteroides* were identified in Moturis and Spring Valley water. *Prevotella*, *Parabacteroides*, *Ruminococcus* (*Blautia*), *Bifidobacterium*, and *Faecalibacterium*, which are all feces-associated bacteria (12–15), were only identified in Moturis water samples analyzed by shotgun metagenomics. Together, these genera represent 6.89% of the

identified bacteria (classified and unclassified) and 23.5% of the classified bacteria. Based on the differential relative abundance of taxa in soil and water from the six sites, these data may be used to inform future efforts toward microbial source tracking.

**Data availability.** The raw sequencing data are available at the NCBI Sequence Read Archive (SRA) under BioProject accession number [PRJNA898587](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA898587). The SRA accession numbers are listed in Table 1.

## ACKNOWLEDGMENT

This work was supported by the National Science Foundation OPUS MCS program to B.J.C. (award number 1950018).

## REFERENCES

1. Riverkeeper, Bronx River Alliance, Quassaick Creek Watershed Alliance, Sarah Lawrence College Center for the Urban River at Beczak. 2022. Riverkeeper Hudson River tributary community science fecal indicator bacteria dataset (v2). HydroShare. <http://www.hydroshare.org/resource/e22138bd77914201af48fce5bfc458f4>.
2. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.
3. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ pre-processor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>.
4. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bio-Bakery 3. *Elife* 10:e65088. <https://doi.org/10.7554/eLife.65088>.
5. Huttenhower Lab. 2016. KneadData (version 0.5.4). Harvard T. H. Chan School of Public Health. <http://huttenhower.sph.harvard.edu/kneaddata>.
6. Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res* 30:42–46. <https://doi.org/10.1093/nar/30.1.42>.
7. Caporaso JG, Ackermann G, Apprill A, Bauer M, Berg-lyons D, Betley J, Fierer N, Fraser L, Fuhrman JA, Gilbert JA, Gormley N. 2018. 16S Illumina amplicon protocol. Earth Microbiome Project. <http://www.earthmicrobiome.org/protocols-and-standards/16s>.
8. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwards CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852–857. <https://doi.org/10.1038/s41587-019-0209-9>.
9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
10. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and Web-based tools. *Nucleic Acids Res* 41:D590–D596. <https://doi.org/10.1093/nar/gks1219>.
11. Robeson MS, Jr, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA. 2021. RESCRIPt: reproducible sequence taxonomy reference database management. *PLoS Comput Biol* 17:e1009581. <https://doi.org/10.1371/journal.pcbi.1009581>.
12. Okabe S, Okayama N, Savichtcheva O, Ito T. 2007. Quantification of host-specific Bacteroides–Prevotella 16S rRNA genetic markers for assessment of fecal pollution in freshwater. *Appl Microbiol Biotechnol* 74:890–901. <https://doi.org/10.1007/s00253-006-0714-x>.
13. Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL. 2013. A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. *Microb Ecol* 65:1011–1023. <https://doi.org/10.1007/s00248-013-0200-9>.
14. McLellan SL, Eren AM. 2014. Discovering new indicators of fecal pollution. *Trends Microbiol* 22:697–706. <https://doi.org/10.1016/j.tim.2014.08.002>.
15. Carrillo M, Estrada E, Hazen TC. 1985. Survival and enumeration of the fecal indicators Bifidobacterium adolescentis and Escherichia coli in a tropical rain forest watershed. *Appl Environ Microbiol* 50:468–476. <https://doi.org/10.1128/aem.50.2.468-476.1985>.