

Wastewater Purification Process: A Metagenomic Analysis

Y. Timmer, F. Menninga and J. Duiker

2025-04-03

Materials and Methods

Three different sampling points were used within the water purification process. The first sample was taken from the digester, and the second and third samples were taken from the lagoons' influent and effluent. The obtained samples consist of 16S rRNA data and were sequenced using a MinION flowcell (FLO-MIN106, pore version: R9.4.1).

For analyzing the sequenced data, we created a SnakeMake (v. 8.27.1) (Mölder et al. (2021)) pipeline with various tools that perform quality control and taxonomic- and functional analysis. For quality control and trimming Fastplong (v. 0.2.2) was used (Chen (2023)), the reads got filtered based on phred score and any reads with an average phred score < 15 were removed. Kraken2 (v. 2.1.2) (Wood, Lu, and Langmead (2019)) was used for the taxonomic classification of the reads, using Greengenes (DeSantis et al. (2006)) as mapping database.

Kraken2 outputs Kraken2 reports, which were used for visualizing taxonomic diversity. With the reports, Krona pie charts were made using Kronatools (v. 2.8.1) (Ondov, Bergman, and Phillippy (2011)) and Sankey charts were generated using Pavian (v. 1.0) (Breitwieser and Salzberg (2020)). Taxonomic diversity was also calculated using the Bray-Curtis dissimilarity (beta diversity). This is done using Kraken Tools (Lu et al. (2022)) with a Python script (beta_diversity.py).

For functional analysis the Kraken2 reports were converted to json formatted .biom files using kraken-biom (v. 1.2.0) (Dabdoub (2016,)) and mapped to a database of metabolic- and ecological pathways using FAPROTAX (v. 1.2.10) (Louca, Parfrey, and Doebeli (2016)).

An overview of the pipeline steps is displayed in fig.1

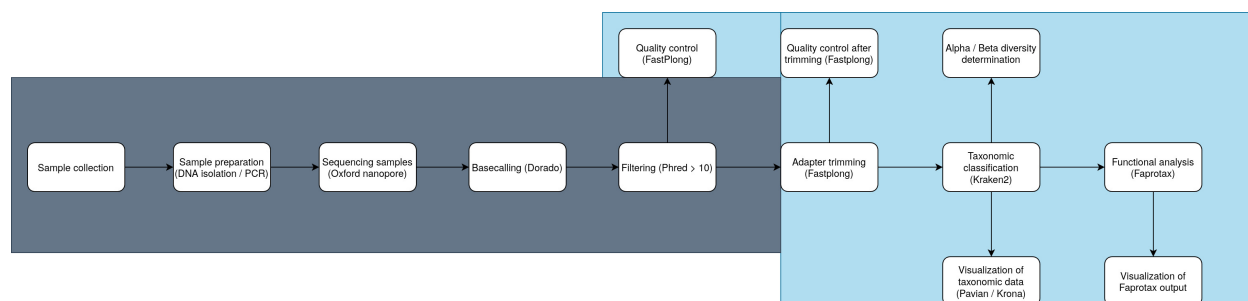


Figure 1: Overview of pipeline steps, including the pre-processing steps that were not performed by us (gray) and the steps that were executed by us (blue).

Results

After running the pipeline the following results were gathered.

Table 1: Overview of read statistics before and after filtering with Fastplong .

	Influent	
Statistic	Before Filtering	After Filtering
Total Reads	20.242K	18.042K
Mean Length	1.426K	1.301K
GC Content	54.45%	51.67%

Table 2: Overview of read statistics before and after filtering with Fastplong out lagoon.

	Effluent	
Statistic	Before Filtering	After Filtering
Total Reads	2.41K	2.16K
Mean Length	1.41K	1.32K
GC Content	53.65%	53.16%

Table 1 shows a decrease in total reads from ~20K to ~18K in the influent. The mean read length is around 1.4K bases and the GC content is 53%. As seen in table 2, The effluent contains 2.16K reads after filtering with Fastplong with a mean length of 1.32K bases. The GC content is 53%.

Taxonomic analysis

Using Kraken2 reports, Krona pie charts were generated. These pie charts show the microbial diversity of the two samples (lagoon in/lagoon out) (fig.1, fig.2). Visibly most of the operational taxonomic units (OTU's), of both samples, consist of cellular organisms and a small percentage of transposons. Most of the cellular organisms consist of the domain Bacteria, the rest consists of a small percentage of Methanobacteriota (archaea). The entirety of OTU's in both samples, in the domain of Bacteria consists of the kingdom Bacillati. In the deeper taxonomic layers, the diversity in OTU's seems to differ in the two samples.

In the lagoon influent (fig. 1), Bacillati divides into mostly Actinomycetes and a smaller percentage Bacillota. The most common species are *Streptomyces noursei* (20%), *Geodermatophilus obscurus* (15%) and *Pseudonocardia nitrificans* (12%). These 3 species make up 47% of total diversity. 14% is classified as 'other root'

In the lagoon effluent (fig. 2), at phylum level, there is a visible divide between Actinomycetes and Bacillota, with the majority of OTU's consisting of Actinomycetes. The most common species are *Mycobacterium kansasii* (22%), *Thermoanaerobacter sp* (20%) and *Streptomyces viridochromogenes* (9%). These 3 species make up 51% of total diversity. 6% is classified as 'other Root'.

The Sankey charts (fig. 3, fig. 4) show only OTU's within the Bacteria domain, across both samples. However the charts display a difference in diversity across mostly phylum level and down. The lagoon influent (fig. 3)

contains a seemingly even distributed number of OTU's in the phyla Bacilli and Clostridia, where the lagoon effluent (fig. 4) contains a much higher number of OTU's in the Clostridia phylum. On order level, Lactobacillales seems a lot more common in the lagoon influent than the effluent. All of the the Lactobacillales OTU's in the lagoon influent seem to belong to the Streptococcus genus.

The beta diversity (Bray-Curtis dissimilarity) resulted in a ratio of 0.89.

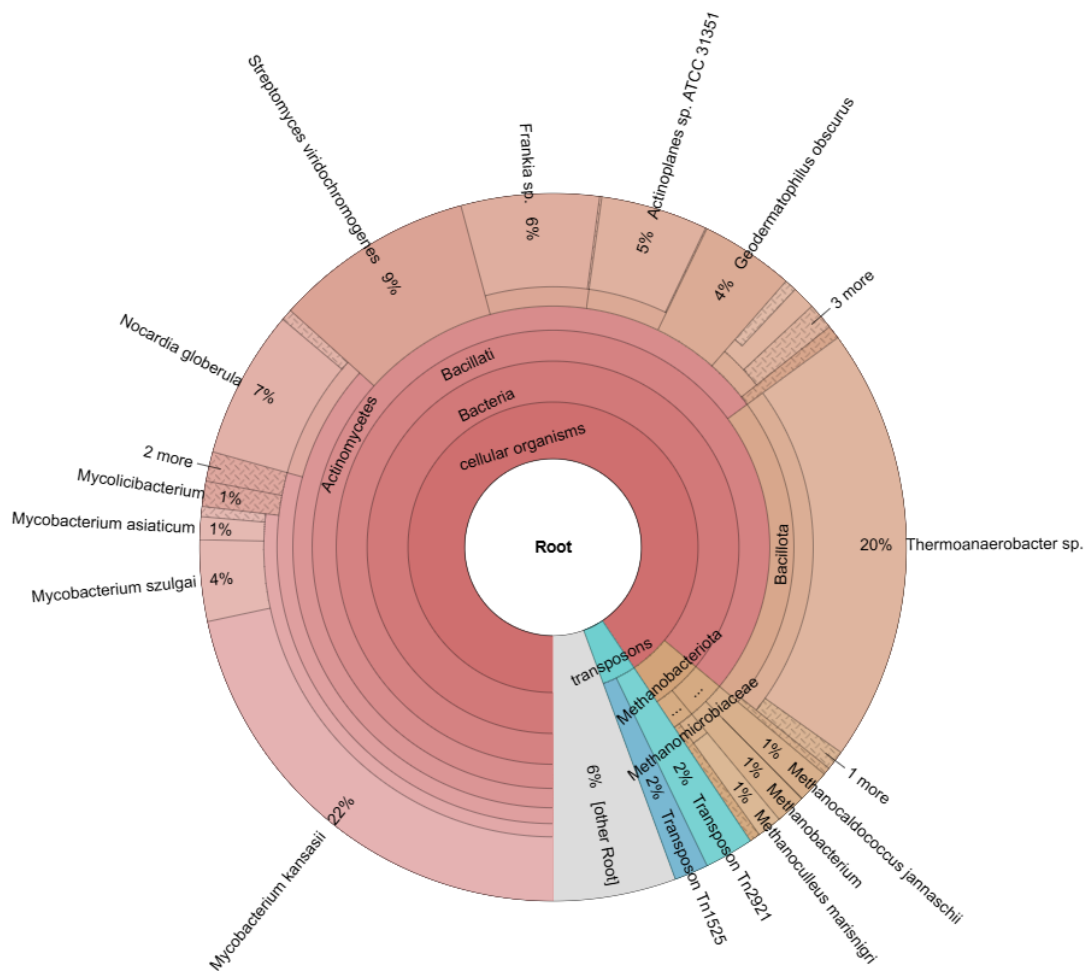


Figure 2: A pie chart showing the microbial diversity of the lagoon influent, generated using Krona.

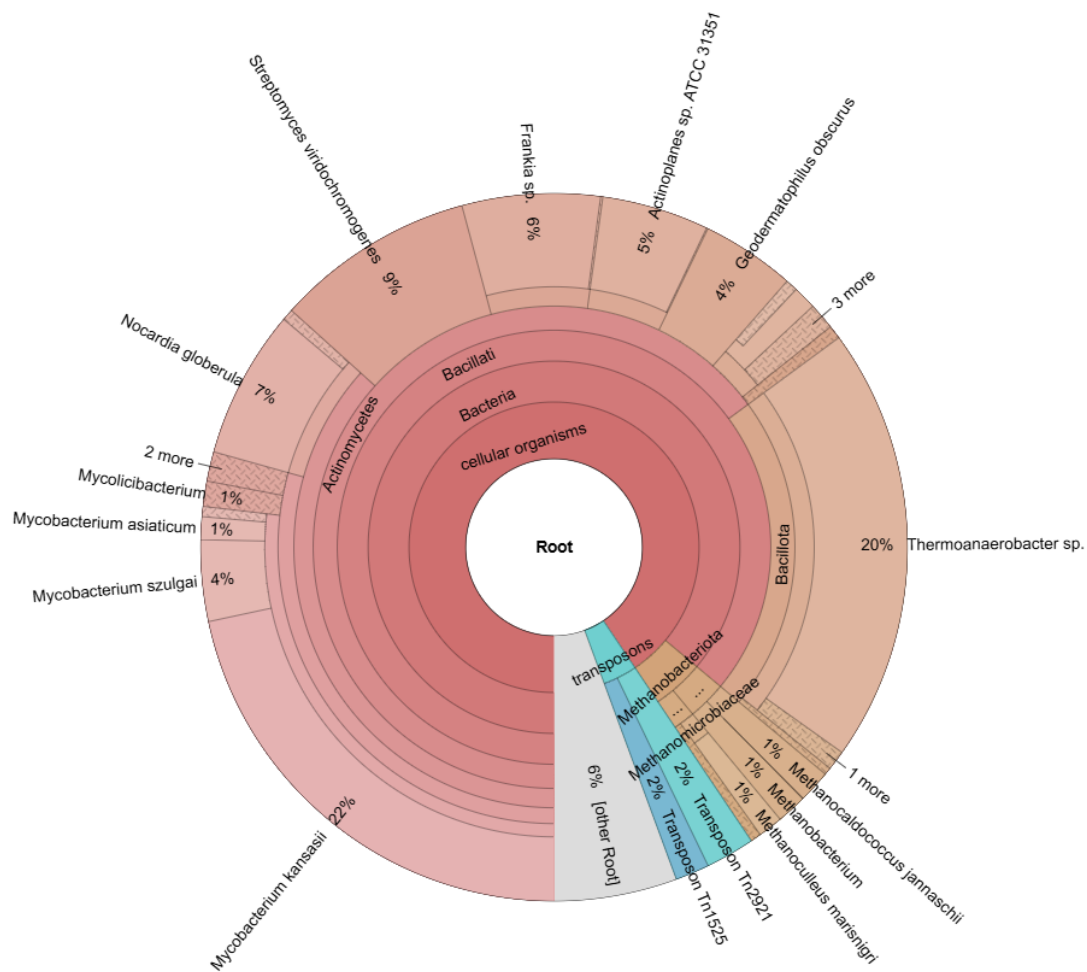


Figure 3: A pie chart showing the microbial diversity of the lagoon effluent, generated using Krona.

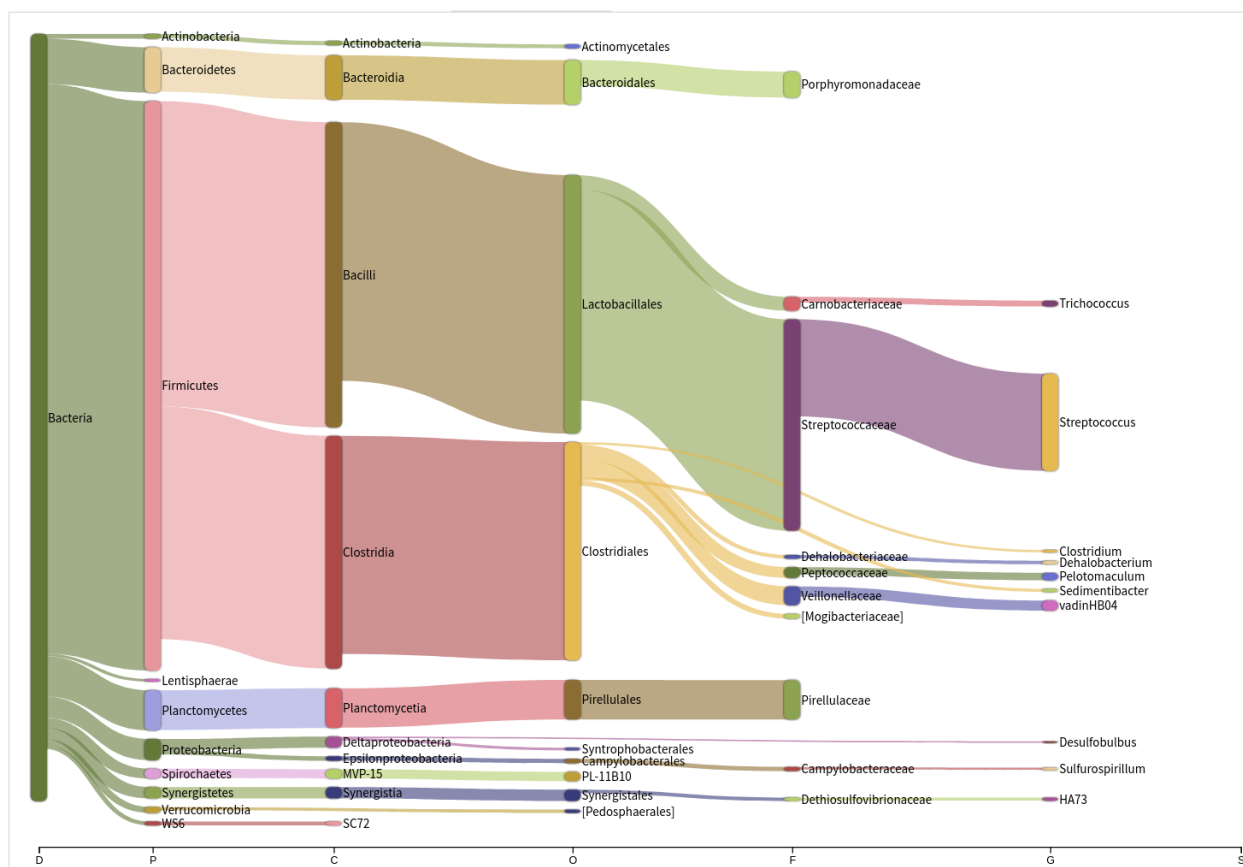


Figure 4: Sankey chart visualizing taxonomic diversity of lagoon influent, generated with Pavian.

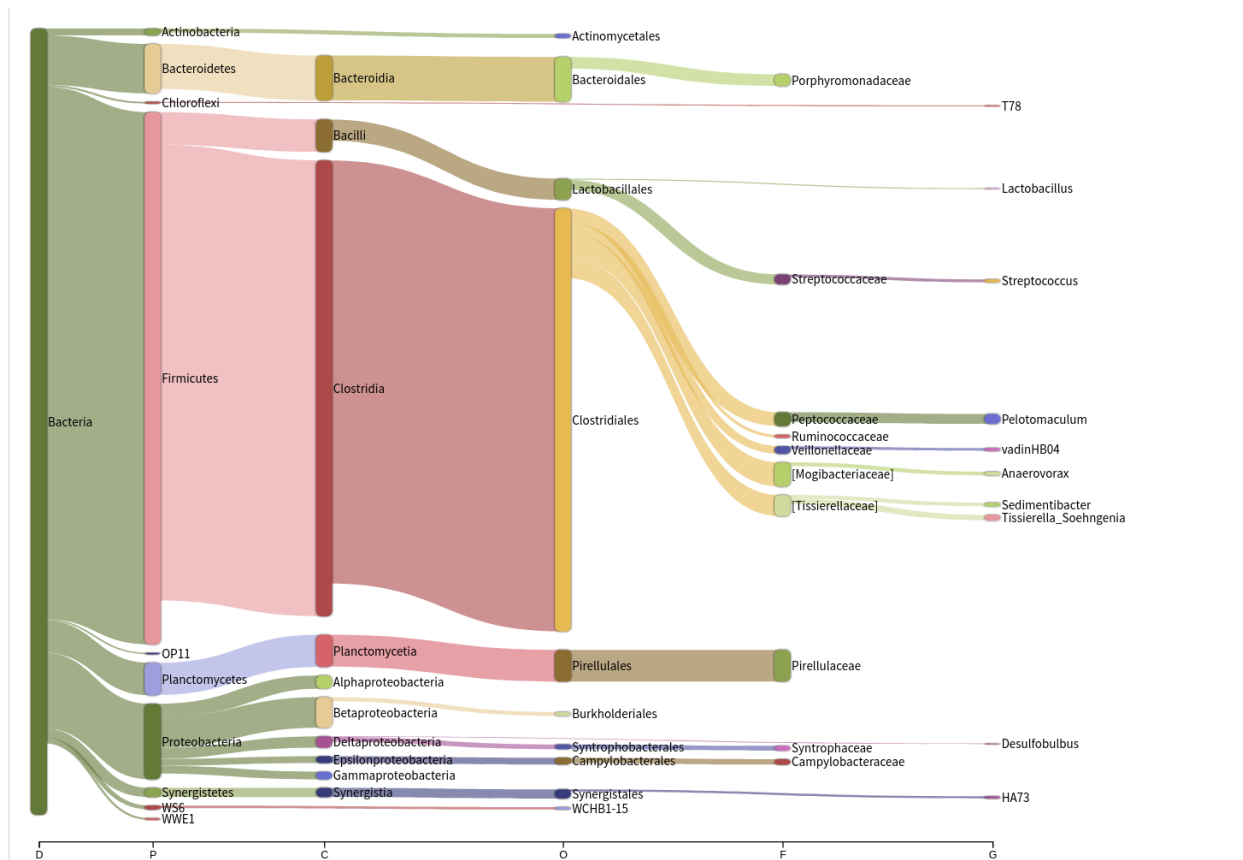


Figure 5: Sankey chart visualizing taxonomic diversity of lagoon effluent, generated with Pavian.

Functional analysis

The classified species were compared to the FAPROTAX database and from here certain (metabolic) pathways were found (fig. 5). The lagoon influent contains OTU's that are known to play a role in the following eight pathways: sulfur respiration, sulfite respiration, sulfate respiration, respiration of sulfur compounds, iron respiration, fermentation, chemoheterotrophy and aromatic compound degeneration. The last three being the pathways with the highest abundance. The lagoon effluent contains OTU's that are known to play a role in the following three pathways: fermentation, chemoheterotrophy and aromatic compound degeneration. These three pathways seem to overlap with the three pathways with the highest abundance of the lagoon influent. Chemoheterotrophy and aromatic compound degeneration seem to have an equal abundance in both samples, while fermentation seems to have a higher abundance in lagoon effluent.

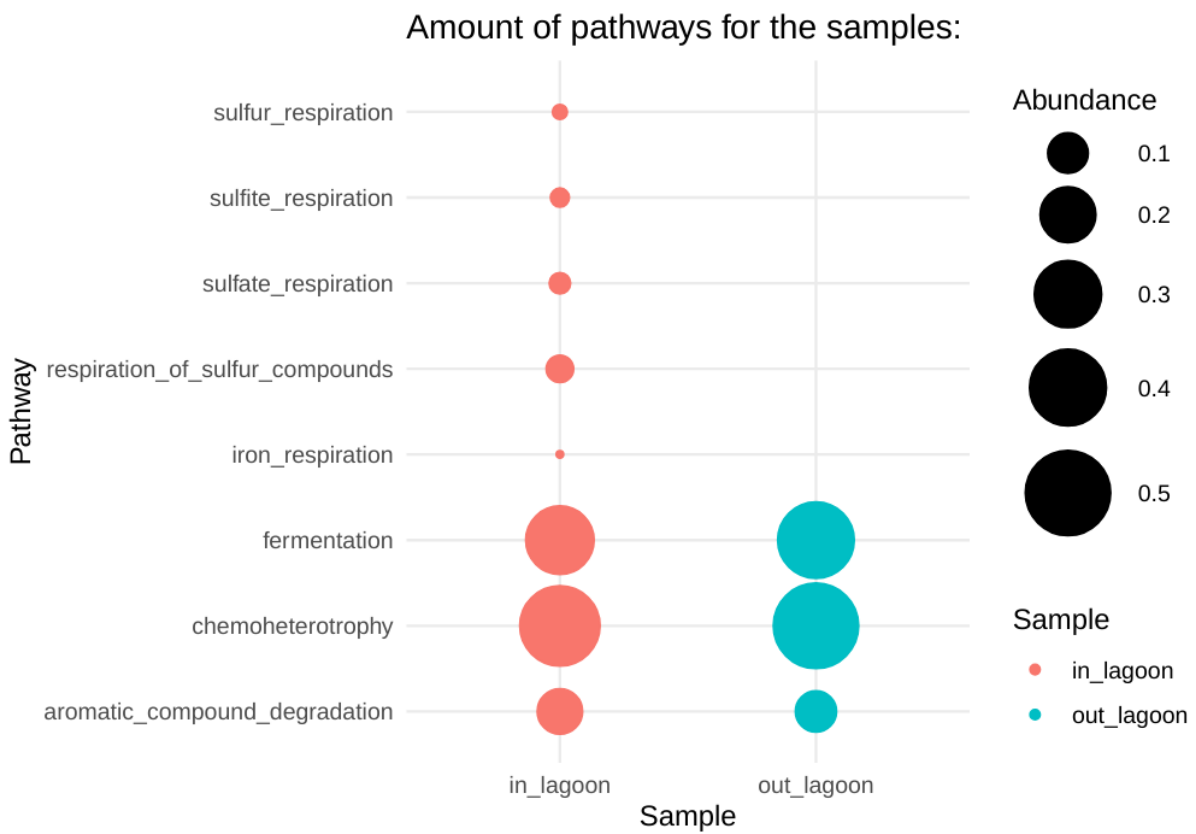


Figure 6: Comparison of read abundance signifying the presence of pathways across samples.

- Breitwieser, Florian P, and Steven L Salzberg. 2020. “Pavian: Interactive Analysis of Metagenomics Data for Microbiome Studies and Pathogen Identification.” *Bioinformatics* 36 (4): 1303–4.
- Chen, Shifu. 2023. “Ultrafast One-Pass FASTQ Data Preprocessing, Quality Control, and Deduplication Using Fastp.” *Imeta* 2 (2): e107.
- Dabdoub, SM. 2016,. “Kraken-Biom: Enabling Interoperative Format Conversion for Kraken Results (Version 1.2) [Software].” <https://github.com/smdabdoub/kraken-biom>.
- DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. 2006. “Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB.” *Applied and Environmental Microbiology* 72 (7): 5069–72.
- Louca, Stilianos, Laura Wegener Parfrey, and Michael Doebeli. 2016. “Decoupling Function and Taxonomy in the Global Ocean Microbiome.” *Science* 353 (6305): 1272–77.
- Lu, Jennifer, Natalia Rincon, Derrick E Wood, Florian P Breitwieser, Christopher Pockrandt, Ben Langmead, Steven L Salzberg, and Martin Steinegger. 2022. “Metagenome Analysis Using the Kraken Software Suite.” *Nature Protocols* 17 (12): 2815–39.
- Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. “Sustainable Data Analysis with Snakemake.” *F1000Research* 10: 33.
- Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. “Interactive Metagenomic Visualization in a Web Browser.” *BMC Bioinformatics* 12 (1): 385. <https://doi.org/10.1186/1471-2105-12-385>.
- Wood, D. E., J. Lu, and B. Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20: 257. <https://doi.org/10.1186/s13059-019-1891-0>.