Investigating Microbial Community in an Anaerobic Reactor for the Treatment of Wastewater from a Molasses-Based Spirit and Yeast Production Factory

Yamila Timmer[1], Floris Menninga[1], Jarno Jacob Duiker[1]
[1]Hanze University of Applied Sciences, Life Sciences, Bioinformatics, Groningen

March 30, 2025

Contents

1 Introduction

Abstract

This study employs metagenomic tools to analyze wastewater from the Afac filtering lagoon in Kenya, which is used by an agrochemical company to treat river water for production processes before release back into the river. The goal is to evaluate the lagoon's effectiveness in removing hazardous microorganisms and assess potential ecological impacts on downstream ecosystems and communities.

Sequencing data from the MinION platform underwent rigorous quality assessment and trimming using Fastplong (Chen et al. 2018), followed by taxonomic classification with Kraken 2 (Wood, Lu, and Langmead 2019). Results were visualized interactively using KronaTools (Ondov, Bergman, and Phillippy 2011) and Pavian (Breitwieser and Salzberg 2020), enabling detailed taxonomic profiling. Microbial functional pathways were analyzed with HUMAnN 3.0 (Beghini et al. 2021), while alpha- and beta-diversity metrics were calculated using a krona-plugin python script.

Our analysis identifies microorganisms persisting through the lagoon's treatment stages, providing insights into filtration efficacy and risks of downstream pollution. The findings will serve as actionable recommendations to optimize the lagoon system, mitigate ecological harm, and safeguard river health for adjacent communities.

Introduction

Countries in northern Africa, the Middle East, Singapore, Maldives, and Australia face water shortages. Making water quite a valuable resource, because of this these countries and regions have practiced reclaiming waste water from factories and sewage. Instead of wasting this water it could help against the big shortages and help the infrastructure stay up and active.

Reclaimed water is used in various places, most known is the landscape irrigation to maintain green living spaces and for agricultural irrigation to produce food (**Hong2020?**). Reclaimed water is also key for important infrastructures like cooling towers that serve electrical power plants. All these things mean that reclaiming this waste water is key to some countries' daily operation.

Reclaiming waste water does come with some risks. The water used in the factories may contain chemicals or microorganisms that could have a negative influence on the health of people or the ecosystem (**Chen2013?**). Many of these factories use filters to take these chemicals and/or microorganisms out of the water so it can return safely into the river or other water source, however not every factory uses the same filter due to costs and/or the infrastructure to make these filters is not available. Thorough filtering is key in reclaiming waste water, because many biological contaminants are disseminated through water, and their occurrence has potential detrimental impacts on public and environmental health (**Hong2020?**).

2 Materials and Methods

2.0.1 Overall Approach

The sequencing data obtained from the MinION platform underwent a quality assessment process. Initially, the sequences were evaluated for quality using `Fastplong` (Chen et al. 2018). After this, `Fastplong` also trimmed the data according to the generated reports. Following this, a secondary quality assessment was conducted using `Fastplong` to ensure improved data quality. The refined dataset was then subjected to taxonomic classification through `Kraken2` (Wood, Lu, and Langmead 2019), a computational tool for microbial classification. The results generated by `Kraken2` were then visualized using `KronaTools` (Ondov, Bergman, and Phillippy 2011) and `Pavian` (Breitwieser and Salzberg 2020), facilitating an interactive and intuitive representation of the taxonomic distribution. Other tools like `HUMAnN 3.0` (Beghini et al. 2021) were used for profiling the abundance of microbial metabolic pathways and other molecular functions from metagenomics data. Furthermore, the outcomes obtained from the `Kraken2` analysis were employed to identify the microorganisms present in the lagoon water, which allowed comparing the different lagoon stages and their microbiomes.

To get more insight into the data, we used two Kraken2 plugins that calculate the alpha- and beta-diversity, including a tool that calculated the Shannon index for alpha diversity:

Shannon index for alpha diversity

$$H = -\sum_{i=1}^{S} p_i \ln p_i$$

where: - H is the Shannon diversity index - S is the total number of species - p_i is the proportion of individuals that belong to species i

This index showed the species diversity in each sample and the distribution of the species in the samples. With all this information, an analysis for each stage of the lagoon has been made.

2.0.2 Data Collection

2.0.2.1 QIAamp DNA Microbiome Kit (50) - 51704 This kit is used for purification and enrichment of bacterial microbiome DNA from swabs (and body fluids). Effective depletion of host DNA during

the purification process maximizes bacterial DNA coverage in NGS analysis and allows for 16rDNA-based microbiome analysis and whole metagenome shotgun sequencing studies.

Procedure

The kit employs spin column technology with a specialized protocol to enrich bacterial microbiome DNA while minimizing host DNA contamination. First, host cells are gently lysed, and their released DNA is enzymatically degraded. Next, bacterial cells are disrupted using optimized mechanical and chemical lysis. The bacterial DNA is then selectively bound to a silica membrane, purified through washes, and finally eluted for analysis.

This method ensures efficient isolation of bacterial DNA from complex samples, reducing host DNA interference.

2.0.2.2 Sequencing

We used Rapid sequencing amplicons - 16S barcoding (SQK-16S024).

Note: The flow cell used in our MinION was not of the highest quality/able to produce a good read amount for our last sample. This resulted in only the Lagoon in and out samples being usable.

The MinION Flow Cell can generate up to 50 Gb of data for sequencing DNA, cDNA or native RNA in real-time. A flowcell is a core sensing unit made up of nanopores, an array of microscaffolds that supports membrane and embedded nanopore. The array keeps the multiple nanopores stable during shipping and usage. Each microscaffold corresponds to its own electrode that is connected to a channel in the sensor array chip. Sensor arrays may be manufactured with any number of channels and an ASIC (Application-Specific Integrated Circuit), with each nanopore channel being controlled and measured individually by the bespoke ASIC. This allows for multiple nanopore experiments to be performed in parallel.

More information is available at: https://nanoporetech.com/platform/technology/flow-cells-and-nanopores

Using MinKNOW, the sequencing was started. To configure the settings needed for MinION sequencing, the user had to go to the kit page in MinKNOW and select the kit used for the library preparation.

The risk of using an old MinION flowcell is that the flowcell can change over time, and the change between each sequencing run is hard to predict.

For the downstream analysis, quality check and trimming were needed to ensure good quality. After using `fastplong`, it was confirmed that our flowcell indeed had problems with the digester samples and did not provide usable reads.

2.0.3 The Metagenomics Pipeline

2.0.3.1 Data Preprocessing and Quality Control

Raw sequencing data was received in FASTQ format and preprocessed using `fastplong`. This specialized version of the fastp tool is optimized for Nanopore long-read data. Adapters and low quality reads (Phred score < 20) were trimmed, and reads shorter than 50 bp were discarded to ensure high-quality data and maintain integrity for downstream analysis. After the initial trim, data quality was reassessed using `fastplong`. It was concluded that no further trimming was needed, allowing the next step to proceed.

2.0.3.2 Identifying Microorganisms Using Taxonomic Classification

The processed FASTQ files were analyzed using `Kraken2` for taxonomic classification of microorganisms. `Kraken2` achieves high accuracy by matching k-mer sequences (nucleotide sequences). The output file contains important information including taxonomy ID, classification status, and the lowest common ancestor list. While taxonomy IDs reveal microorganism identities, manually searching each ID would be inefficient. This challenge is addressed in the visualization step.

2.0.3.3 Visualization of Kraken2 Results To visualize the taxonomic classification results from `Kraken2`, we used multiple tools:

- `KronaTools` created Krona plots showing all taxonomic levels from superkingdom to family level, with associated abundances based on identified spectra. These plots provide an intuitive representation of the findings.
- `Pavian` provided an interactive browser application for analyzing and visualizing metagenomics classification results, including:
  - Taxonomic classification charts
  - Sankey diagrams
  - An alignment viewer to validate genome matches

2.0.3.4 Other Visualizations `HUMAnN 3.0` profiled the abundance of microbial metabolic pathways and molecular functions, revealing the metabolic potential of the lagoon's microbial community. This helped answer the question: "What are the microorganisms in the lagoon doing or capable of doing?"

`QIIME2` calculated the Shannon index for alpha diversity, showing species diversity and distribution in each sample. These distributions enabled us to build microbiome structures for each lagoon stage, though specific research questions about the lagoon remain to be clarified. # Results

[Your results content here]

3 Discussion

[Your discussion content here]

4 Conclusion

[Your conclusion content here]

5 Funding

[Details of funding sources]

6 Acknowledgements

These should be included at the end of the text and not in footnotes. Please ensure you acknowledge all sources of funding.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Mìguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with BioBakery 3." *eLife* 10: e65088. https://doi.org/10.7554/eLife.65088.

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. https://doi.org/10.1093/bioinformatics/btz715.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor." *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

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.