Compositional and Functional Trends in Activated Sludge Bacterial Communities
Project Deliverable (Short summary)
Farris Tedder, Ruben Lancaster, Leyla Cufurovic
3/15/2024

This document is meant to update Dr. Coats on the current state of the project and act as a quickly digested guide as we move forward with preparing to publish.

Background

Activated sludge (AS) is a mixture of microbes that play an important part in wastewater treatment (WWT). Living bacteria within AS remove polluting nutrients such as phosphorus and nitrogen before effluent is released back to the water cycle. This study focused on the taxonomic composition of these bacterial communities and their associated functional potential, suggesting the roles they may play in AS. AS was sampled from various wastewater treatment setups: from a full-scale operational waste water treatment plant in Moscow, ID, to experimental sequencing batch reactors used for PHA (polyhydroxyalkanoates, a bioplastic precursor) production, to EBPR (Enhanced Biological Phosphorus Removal) and nitrogen removal systems.

The data span a range of years, and samples were collected from a variety of setups. Despite this variety, these systems can be grouped by their intended specialized function: to remove phosphorus (P), to remove phosphorus and nitrogen (PN), or to enhance PHA production (PHA). Because much of our analysis focused on the functional potential of these systems, we decided to group systems by intended function to investigate whether this affected functional potential inferred from bacterial taxonomy.

Methods

Sample collection and preparation: Activated sludge samples were collected by the Coats lab from a variety of system setups from 2014-2018 (see Supplementary Table 1, **Illumina_Reads_updated_1-12-24.csv**) and amplicon sequencing was performed by the IIDS Genomics and Bioinformatics Resources Core at the University of Idaho. Forward and reverse reads were obtained from Illumina sequencing of the V4 region of the 16s rRNA gene.

Sequence processing: Raw sequencing reads were processed using the DADA2[1] pipeline implemented in R. Quality filtering and trimming were performed on the raw reads to remove low-quality bases and filter out chimeric sequences. Error rates were estimated and reads were denoised to obtain amplicon sequencing variants (ASVs).

Taxonomic assignment: Taxonomic assignment of ASVs was conducted using the SILVA[2,3,4] database. Species-level assignment was performed and used to create a phyloseq[5] object for downstream microbial community analysis.

Data preprocessing: ASVs found in only one sample were removed along with sequences identified as mitochondria (n = 311), chloroplasts (n = 39), or eukaryotes (n = 1). Rarefaction curves were run for all samples to assess the effect of sampling effort on species richness. Samples with sampling effort below the asymptote of the rarefaction curve were discarded. This means 4 samples with below 10,000 ASVs were discarded. After preprocessing, the dataset included 84 unique samples and 6,037 unique taxa. Normalization and variance stabilization was performed using DESeq2[6] prior to analyzing beta diversity.

Data Analysis: Various analyses were performed using the phyloseq package in R.

Alpha diversity metrics including richness, Shannon's diversity index, and Simpson's diversity index were calculated to assess microbial diversity within samples using the vegan package. For further comparisons of alpha diversity, Shannon's diversity index was chosen for consistency because it is the primary measure of richness and evenness used in prior work from the Coats lab[7,8].

Beta diversity was analyzed using non-metric multidimensional scaling (NMDS) with Bray-Curtis distances to visualize community composition differences among samples using the phyloseq and vegan[9] packages. Differential abundance analysis (ANCOM-BC) was conducted to identify taxa that significantly differed between sample groups using the ANCOMBC2[10] package.

Phylogenetic analysis: Sequence alignment and phylogenetic tree construction were performed using the DECIPHER[11] package in R. Aligned sequences were used to build a maximum likelihood phylogenetic tree, which was integrated into the phyloseq object to use for downstream analysis.

Visualization: The ggplot2[12], plotly[13], and MicEco[14] packages were used for visualization of alpha diversity (Shannon's index) and beta diversity (microbial abundance graphs, NMDS plots, and Euler diagrams). fantaxtic[15], an add-on to phyloseq, was used for visualizing the abundances of taxons by taxonomic rank. rstatix[16] was used for visualizing significance bars on plots.

Statistical analysis: Statistical significance testing was conducted using ANOVA to compare alpha diversity metrics between groups and PERMANOVA[17] with PERMDISP[18] (as the adonis function in the vegan package) for assessing beta diversity. ANCOM-BC[19] (Analysis of Compositions of Microbiomes with Bias Correction) was performed to assess globally differentially abundant taxa.

Data availability: All code and most data (metadata, taxonomic assignment, and ASV counts file) are available at: https://github.com/Farrisdt/SludgeCommunityAnalysis. Raw sequence files are not hosted on github but are available upon request.

Results

Alpha diversity analysis

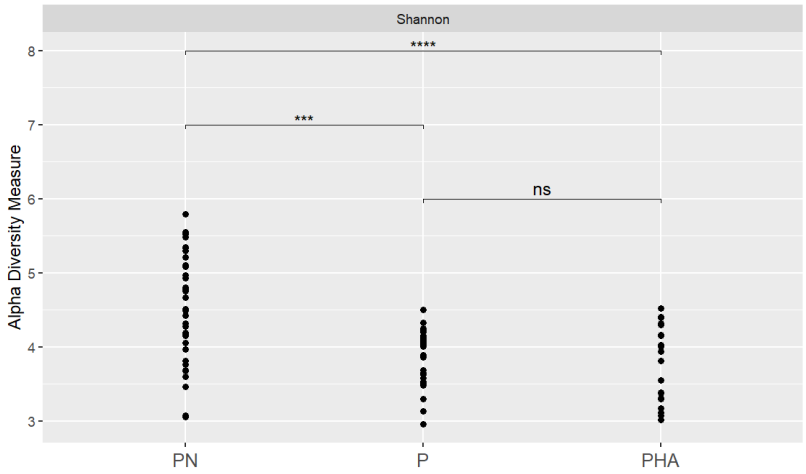


Figure 1. PN systems have significantly different alpha diversity measures (Shannon's diversity index) from P systems and PHA bioreactors (ANOVA with Tukey HSD, $p < 0.00001$). PHA bioplastics reactors and P systems do not have significantly different alpha diversity measures (ANOVA with Tukey HSD, $p = 0.952$).

Alpha diversity analysis using Shannon's diversity index suggests there is greater evenness and richness of the structure of bacterial communities in PN systems (Figure 1). PN systems were entirely fed with raw wastewater instead of synthetic wastewater, in addition to operating differently than the other systems with abiotic processes intended to increase the rate of nitrogen cycling. All of these factors may together explain why PN systems trended higher on the Shannon's diversity measure, but more analysis is needed to identify common ground between samples with the highest Shannon indices that may not have been captured in our metadata.

Beta diversity analysis

Beta diversity analysis revealed differences in the structure of bacterial communities along several axes, including intended system function, timepoint, inflow media type, and system size.

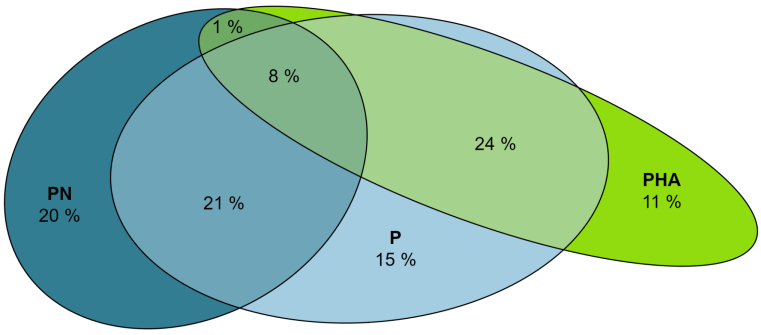


Figure 2. Euler diagram of taxa shared between PN, P, and PHA systems.

A Euler diagram (Figure 2) illustrates that each system has taxa unique to it (PN: 20%, P: 15%, PHA: 11% of total unique taxa), and systems also have shared taxa. 8% of taxa identified appear in all three systems out of the total taxa identified (n = 6,037). The least overlap exists between PHA and PN systems. Euler diagrams split by inflow media type (synthetic wastewater, real wastewater, or fermented dairy manure) and operational setup (full-sized facility, scale-model facility, or bench top sequence batch reactor) followed similar patterns (figures not shown).

We decided to tackle this problem of analyzing and visualizing these highly diverse communities by finding the most populous orders of bacteria. Orders were chosen as this allowed the highest level of taxonomic resolution achievable with the least amount of missing data. Of the top ten orders identified, half of them were globally differentially abundant, suggesting that even among the most populous groups of bacteria, their abundance may change across different environments (Figure 3). This is also reflective of differences of overall bacterial community structure between system setups.

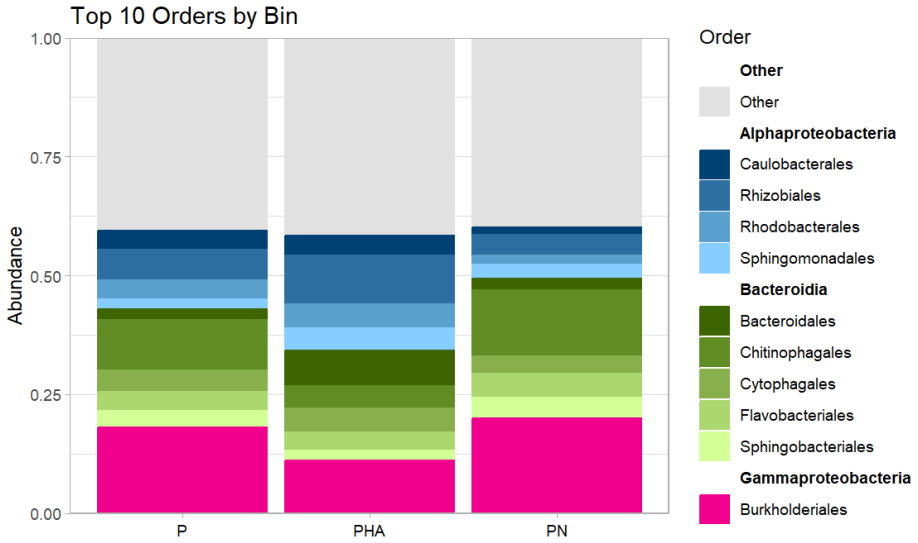


Figure 3. Many predominant orders are differentially abundant in AS. Shown are proportions of the 10 most represented taxonomic orders out of 114 total taxonomic orders, by system. Caulobacterales, Rhodobacterales, Chitinophagales, Cytophagales, and Sphingobacteriales are differentially abundant between all three systems (ANCOM-BC, $p < 0.001$).

These overall community differences were also supported by visualization with non-metric multidimensional scaling with Bray-Curtis distances (Figure 4).

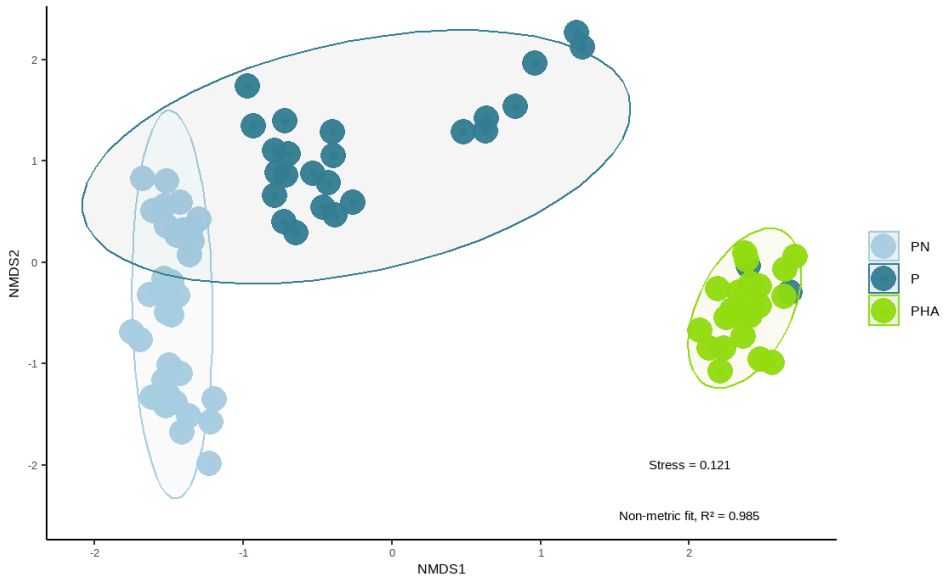


Figure 4. Samples cluster by system. Non-metric multidimensional scaling ordination with Bray-Curtis distances. Samples are points and axes are arbitrary. Centroids of clusters are significantly different (PERMANOVA, $p < 0.001$), and the spread of clusters (group dispersion) is also significantly different (PERMDISP, $p = 0.04$).

Functional Analysis

FAPROTAX[20] is a database of bacterial functional tags associated in literature. It has a python script that returns a report of each function found with the ASVs associated with it. This allows us to get a count of both the functions available in the community and the amount of bacteria associated with them.

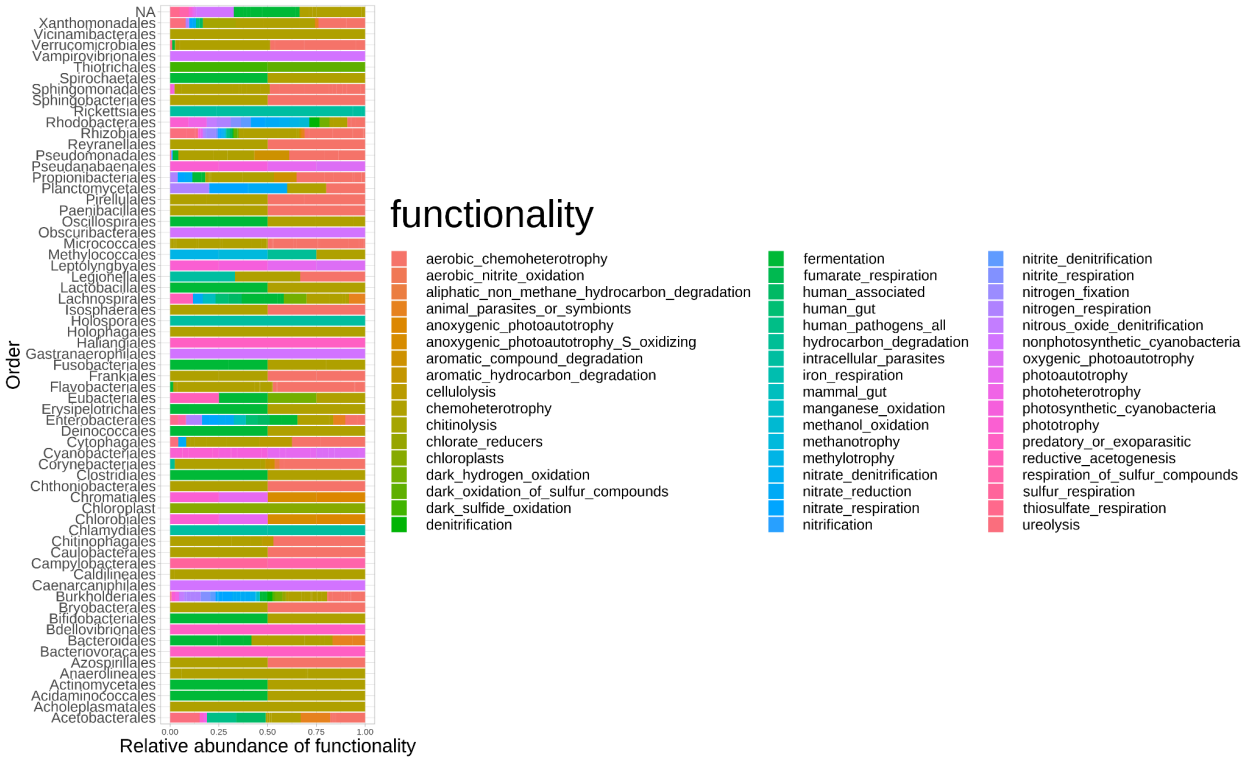


Figure 5. Relative abundance of functional tags of full dataset. Includes all 65 orders and 51 associated functions.

This figure shows all orders (n=65) and their associated functions (n=51). In order to better represent the data we filtered to only the 8 most abundant orders in the data. This reduced the functions represented (n = 36). This was also a way to remove any non-bacterial contaminants, such as chloroplasts, from the analysis dataset.

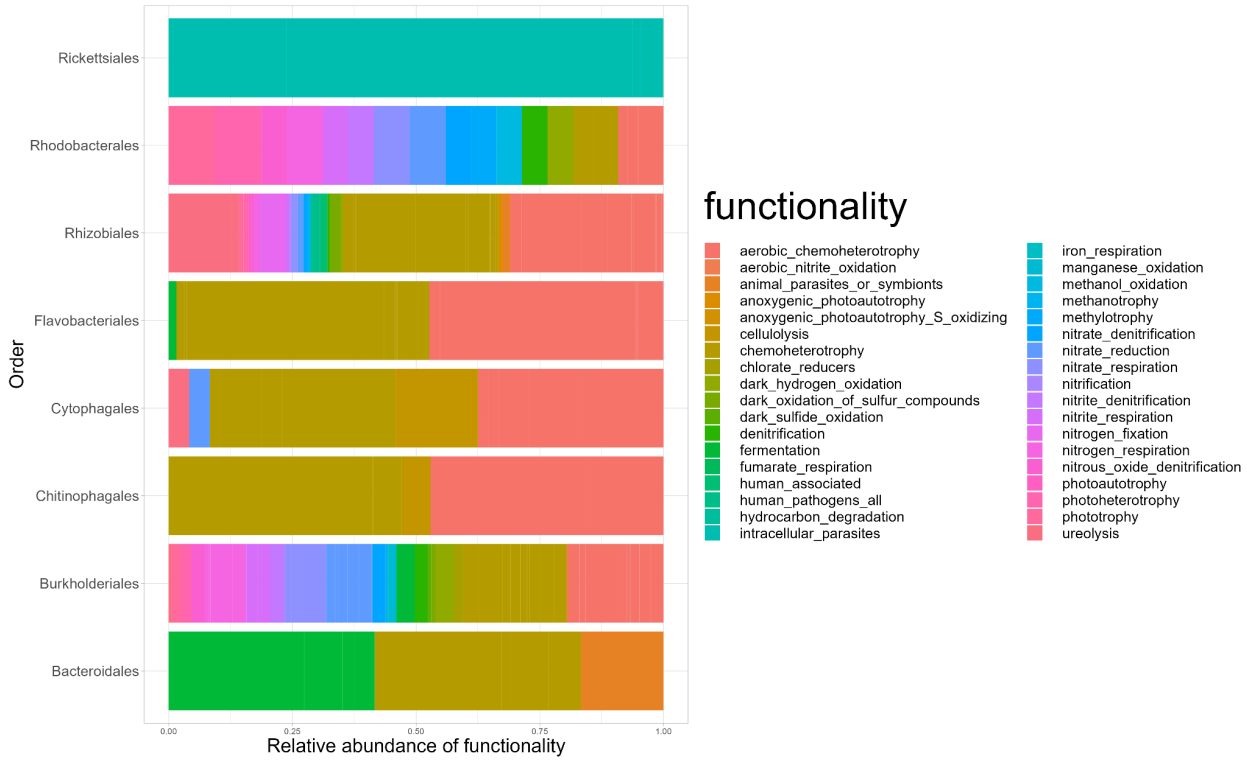
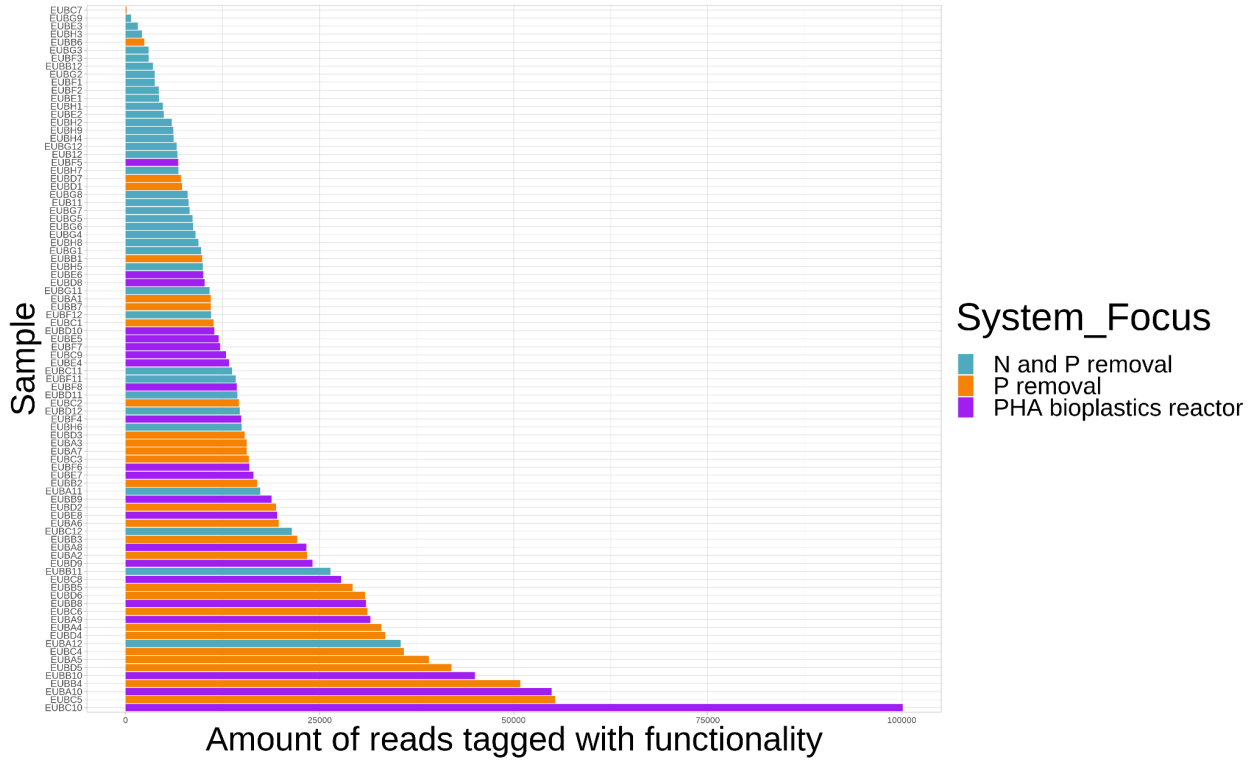


Figure 6. Relative abundance of functional tags of 8 most populous orders. Includes all 36 functions associated with these orders.

The final filtering done was to remove redundancies in the function. Some of the functions had the same representation by different names, such as aerobic_chemoheterotrophy and chemoheterotrophs. All aerobic chemoheterotrophs were represented in chemoheterotroph, making the group appear twice as large as was accurate. A more detailed report on what was removed and what function it was overlapping with is in the R file (section “relative functionality in top orders with data correction”).

A.



B.

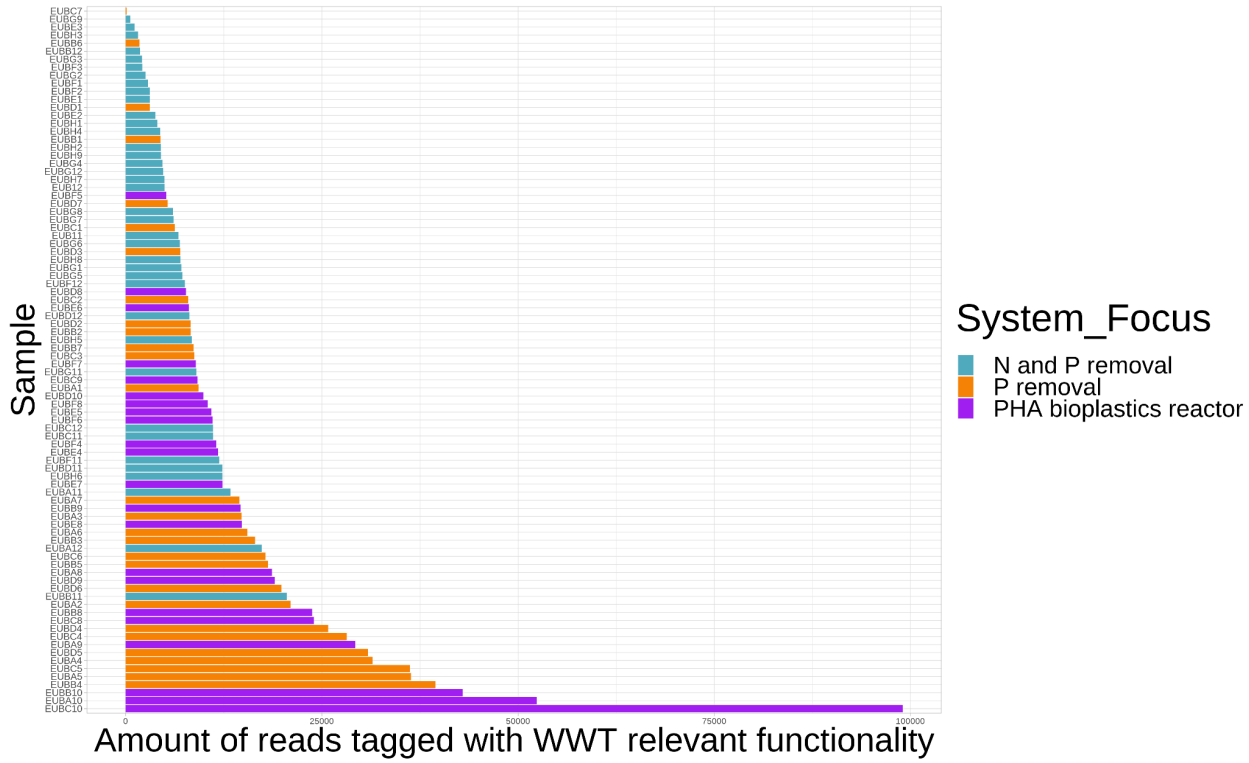


Figure 7. Counts of all functional tags associated with all samples. Samples are colored by the type of system they came from. (a) All functions and (b) only functions relevant to waste water

system processing. Sample names are not relevant, and just shown to mark the data is coming from a variety of samples.

Above are graphs of the number of functional tags per sample, colored by what type of system they were taken from. The bottom graph is only looking at functions deemed important to waste water system functioning. The specific axis are less important than the proportions between bars, which give us an understanding of the general distribution of the data.

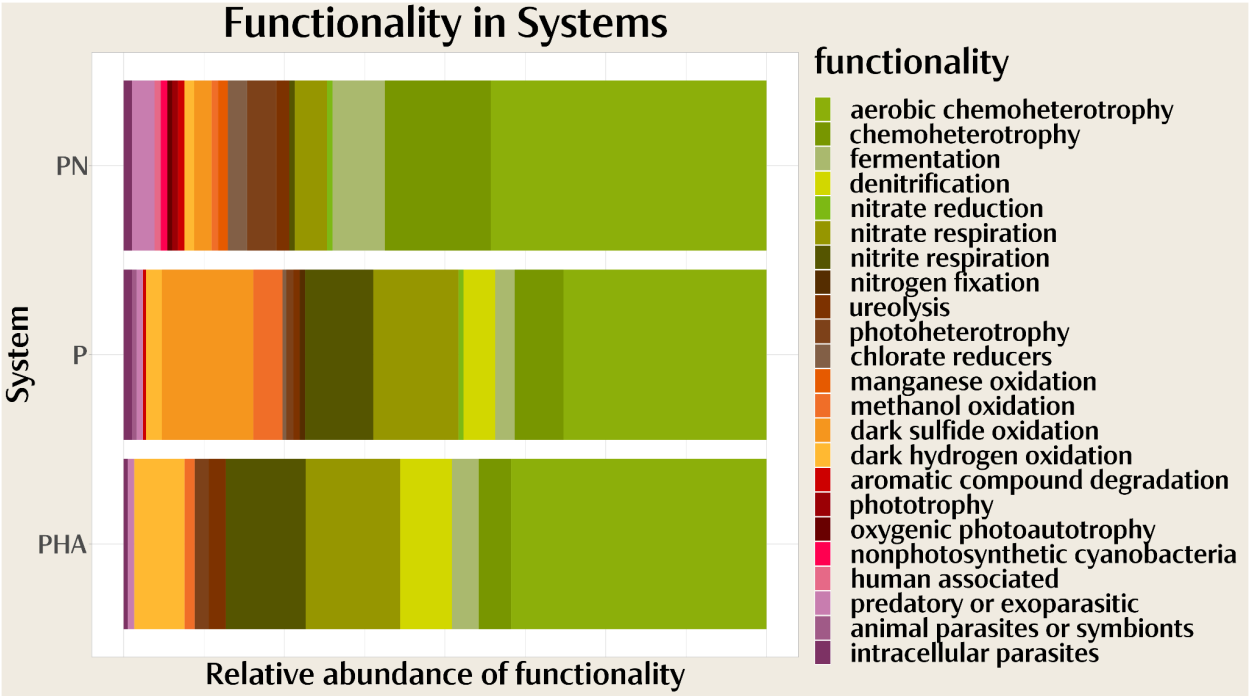


Figure 8. Relative abundance of functionality tags of all bacterial communities, organized by system type. All green functionalities (top 7) are known to be associated with waste water processing.

This graph is made with data after all filtering is complete. In this graph the green functionalities (top 7) are directly related to waste water system functioning, while the other colors have no noted relation. The most important point to take from this is that the most highly represented

functional tags in the communities are ones related to waste water functioning.

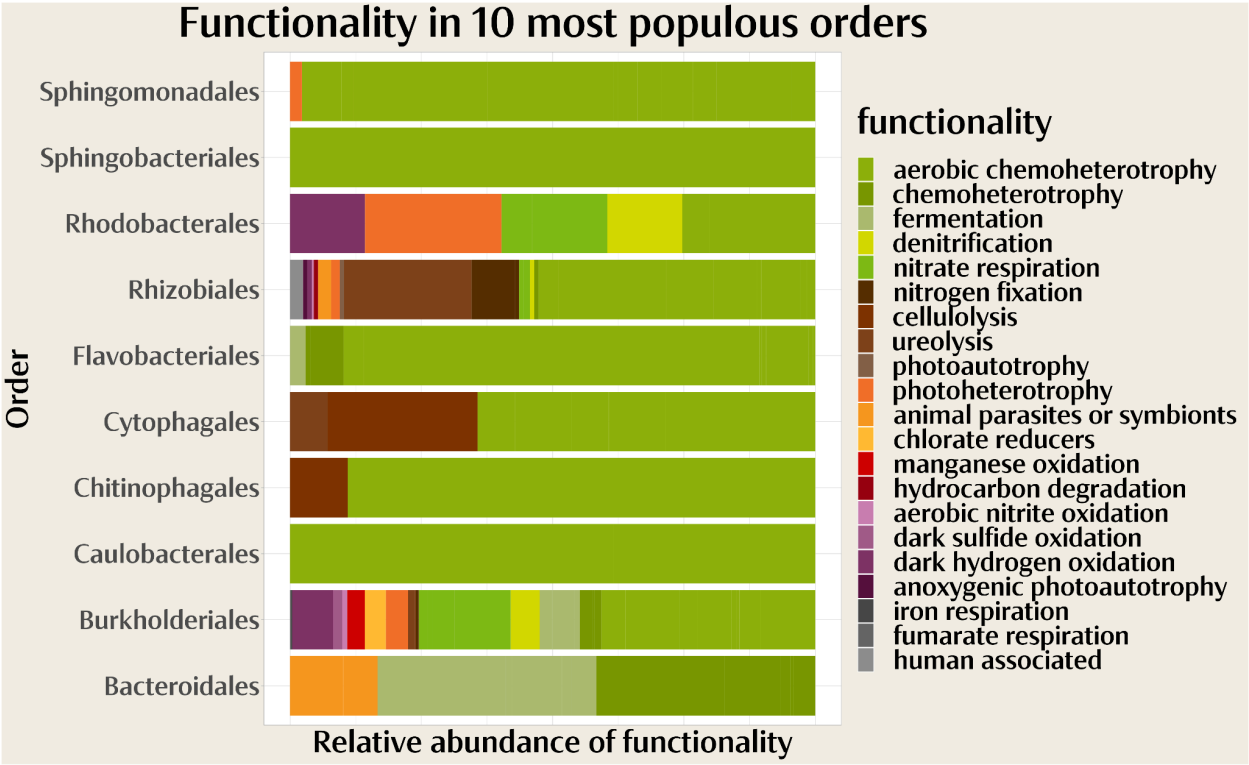
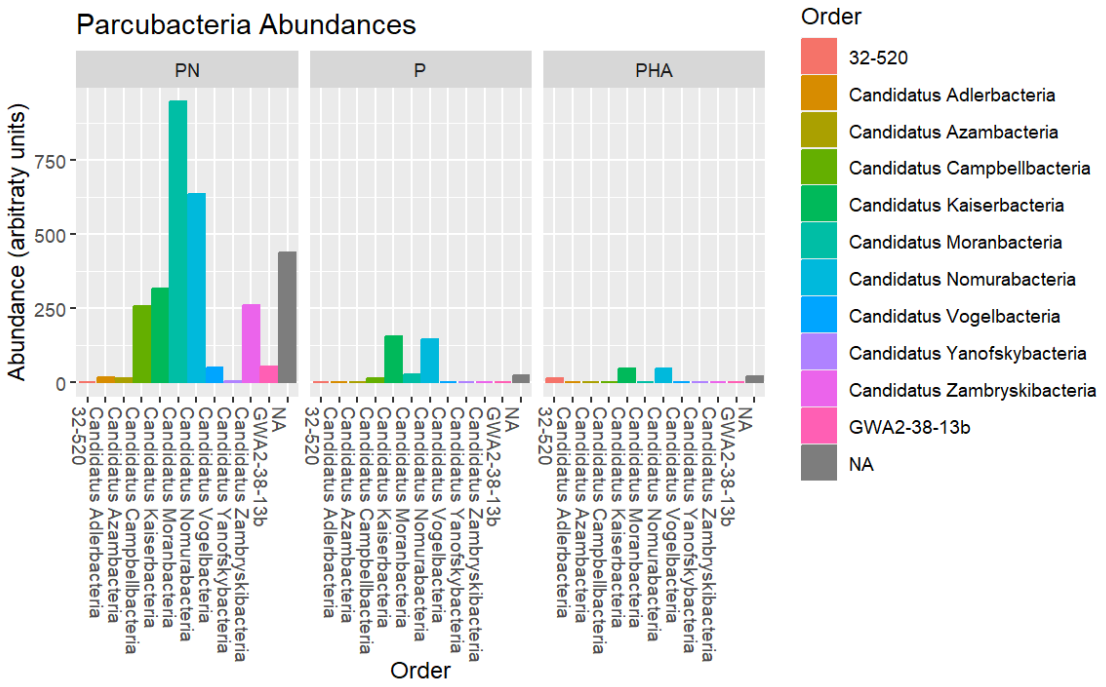


Figure 9. Relative abundance of functionality of the top 10 most represented orders. All green functionalities (top 5) are known to be associated with waste water processing.

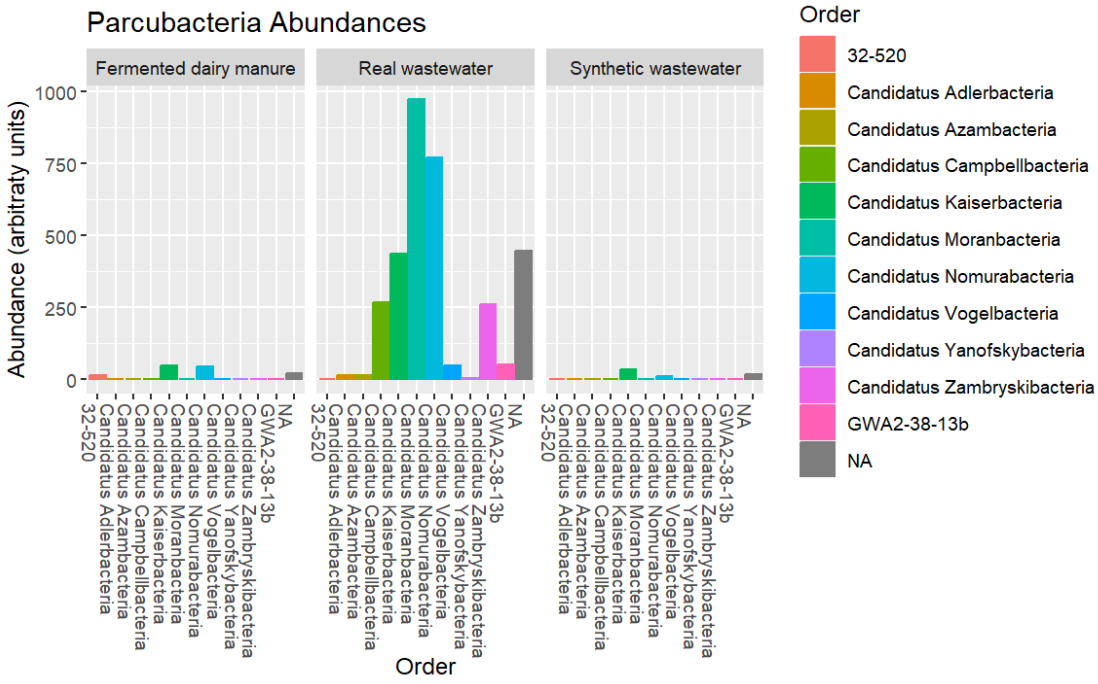
This graph shows the functional tags of the top 10 most populous orders using the same color scheme as before. Here we can see strong differences between the functions of the different orders. That being said, once again all orders have a strong presence of waste water related functional tags. This hints that there may be some functional redundancy occurring between these orders. However, this is tempered by the fact that chemoheterotrophy is a broad category with a lot of research behind it. It is also possible that the chemoheterotrophy we see is actually unrelated to waste water functioning, due to this broad definition. A weakness of this form of analysis is that it is heavily influenced by what the general scientific community finds interesting, making it susceptible to error from low representation of both taxonomy and functional tags. As such, it is important not to draw firm conclusions from this analysis and instead use it as a jumping off point for further, more targeted, research. Further research would be needed to confirm this high presence of waste water function.

Parcubacteria-specific analysis

Efforts were made at the beginning of the project to investigate Parcubacteria specifically. Beta diversity analysis using NMDS proved difficult due to the sparsity of Parcubacteria differences between samples. Presence-absence and abundance testing was possible, however, and revealed that PN systems and systems with real wastewater had a greater amount and variety of Parcubacteria (Figure 10).



A.



B.

Figure 10. Total abundance (counts) in arbitrary units of orders belonging to class Parcubacteria among (**A**) PN, P, and PHA system setups and (**B**) media inflow types (fermented dairy manure, real wastewater, or synthetic wastewater).

This is an interesting observation that could be explored further, but as described below for the purposes of this project it was not ideal to work with Parcubacteria only.

Parcubacteria was not tagged for function through FAPROTAX. We believe this to be due to a lack of research on this taxonomic Class. We previously were presenting communities in which parcubacteria were present to see if the function of parcubacteria communities differed from those without parcubacteria, but found a lack of validity in those studies so chose to omit them. This is a good demonstration on how studying a greater variety of organisms is benificial to the community at large.

Analysis / Conclusions

These results support the conclusion that the taxonomic makeup of bacterial communities differs between system setups.

The majority of functional potential identified in these communities was related to wastewater treatment. Of the top bacterial orders surveyed, some have a wide variety of functional potential, while some are annotated with a single function—suggesting either that some orders have the potential to fill more ecological roles, or that they are better annotated. This supports the hypothesis that there is functional redundancy occurring in these systems. Functional redundancy is an ecological state in which multiple taxa perform the same function; this is important in WWT to ensure consistent processing across different scales, timepoints, and system configurations.

Future work would involve correlating system efficiency data with functional potential and taxonomy to search for bacterial taxa with a high individual effect on system performance.

References

1. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). “DADA2: High-resolution sample inference from Illumina amplicon data.” _Nature Methods_, *13*, 581-583. doi:10.1038/nmeth.3869
2. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.
3. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucl. Acids Res. 42:D643-D648
4. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. J. Biotechnol.
5. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. Paul J. McMurdie and Susan Holmes (2013) PLoS ONE 8(4):e61217.
6. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

7. Coats, E. R., Brinkman, C. K., & Lee, S. (2017). Characterizing and contrasting the microbial ecology of laboratory and full-scale EBPR systems cultured on synthetic and real wastewaters. *Water Research*, *108*, 124–136. https://doi.org/10.1016/j.watres.2016.10.069
8. Coats, E. R., Appel, F. J., Guho, N., Brinkman, C. K., & Mellin, J. (2023). Interrogating the performance and microbial ecology of an enhanced biological phosphorus removal/post-anoxic denitrification process at bench and pilot scales. Water Environment Research, 95(4), e10852. https://doi.org/10.1002/wer.10852
9. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter Braak C, Weedon J (2022). _vegan: Community Ecology Package_. R package version 2.6-4, <https://CRAN.R-project.org/package=vegan>.
11. Wright ES (2016). "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." The R Journal, 8(1), 352-359.
12. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
13. Sievert C (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. ISBN 9781138331457, https://plotly-r.com.
14. Russel J (2023). _MicEco: Various functions for microbial community data_. R package version 0.9.19.
15. Teunisse, G. M. (2022). Fantaxtic - Nested Bar Plots for Phyloseq Data (Version 2.0.1) [Computer software]. https://github.com/gmteunisse/Fantaxtic
16. Kassambara A (2023). _rstatix: Pipe-Friendly Framework for Basic Statistical Tests_. R package version 0.7.2, <https://CRAN.R-project.org/package=rstatix>.
17. Anderson, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online* (pp. 1–15). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat07841
18. Anderson, M.J. 2005. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62:245-253.
19. Lin, H., Peddada, S.D. Analysis of compositions of microbiomes with bias correction. *Nat Commun* 11, 3514 (2020). https://doi.org/10.1038/s41467-020-17041-7
20. Louca, S., Parfrey, L.W., Doebeli, M. (2016) - Decoupling function and taxonomy in the global ocean microbiome. Science 353:1272-1277