Microbial Communities in Alaskan Permafrost: Assessing the Compatibility of Genomic Data

Nora Fried

Abstract

Arctic permafrost soils store vast amounts of organic carbon and are experiencing accelerated thaw under climate warming. This feedback loop has significant implications for global nutrient cycling. Microbial community composition and activity mediate the decomposition of previously frozen soil organic matter, releasing greenhouse gases such as CO_2 and CH_4, yet the relative influence of primary landscape drivers, such as glacial drift and sample depth, remains unclear. Here, I re-analyze soils from D.V. Bakke's master's thesis, assessing two new DNA datasets (RunA and RunB) produced by the re-sequencing of soil samples, with the aim of addressing prior methodological issues that limited D.V. Bakke's analysis. The primary question driving this analysis is whether or not the new DNA data from RunA and RunB can be reliably combined for downstream analyses. Data processing involved data cleaning, rarefaction to standardize sequencing depth, ordination and diversity analyses. This initial assessment exposed a variety of sampling issues within the dataset and provides guidance for next steps in analyzing Bakke's Alaskan permafrost soil samples.

Introduction

The Arctic is warming nearly four times faster than the global average, driving rapid environmental change across this sensitive ecosystem (Rantanen et al. (2022)). One of the most significant consequences of warming is the thawing of permafrost, perennially frozen soils that have accumulated large reservoirs of soil organic carbons over thousands of years. Although permafrost regions cover only ~15% of the global land surface, they store more than twice as much carbon as the atmosphere (E. a. G. Schuur et al. (2015), E. A. G. Schuur et al. (2022)). As permafrost thaws, these carbon pools are destabilized, creating a positive feedback to global climate change through the release of greenhouse gases (GHGs).

Microbes are central to this process of organic carbon degredation: as previously frozen organic matter becomes available, soil microbes rapidly metabolize it, releasing carbon dioxide (CO_2) and methane (CH_4) into the atmosphere (Mack et al. (2004)). The composition and activity

of microbial communities therefore directly influence the rate and magnitude of permafrost carbon cycling. Yet, microbial diversity across Arctic landscapes is highly heterogeneous, and the factors driving this variation remain poorly understood. In particular, geographic environmental gradients such as glacial drift and sample depth are likely to shape microbial community composition and diversity, but their relative contributions have not been systematically evaluated (Bakke (2024)).

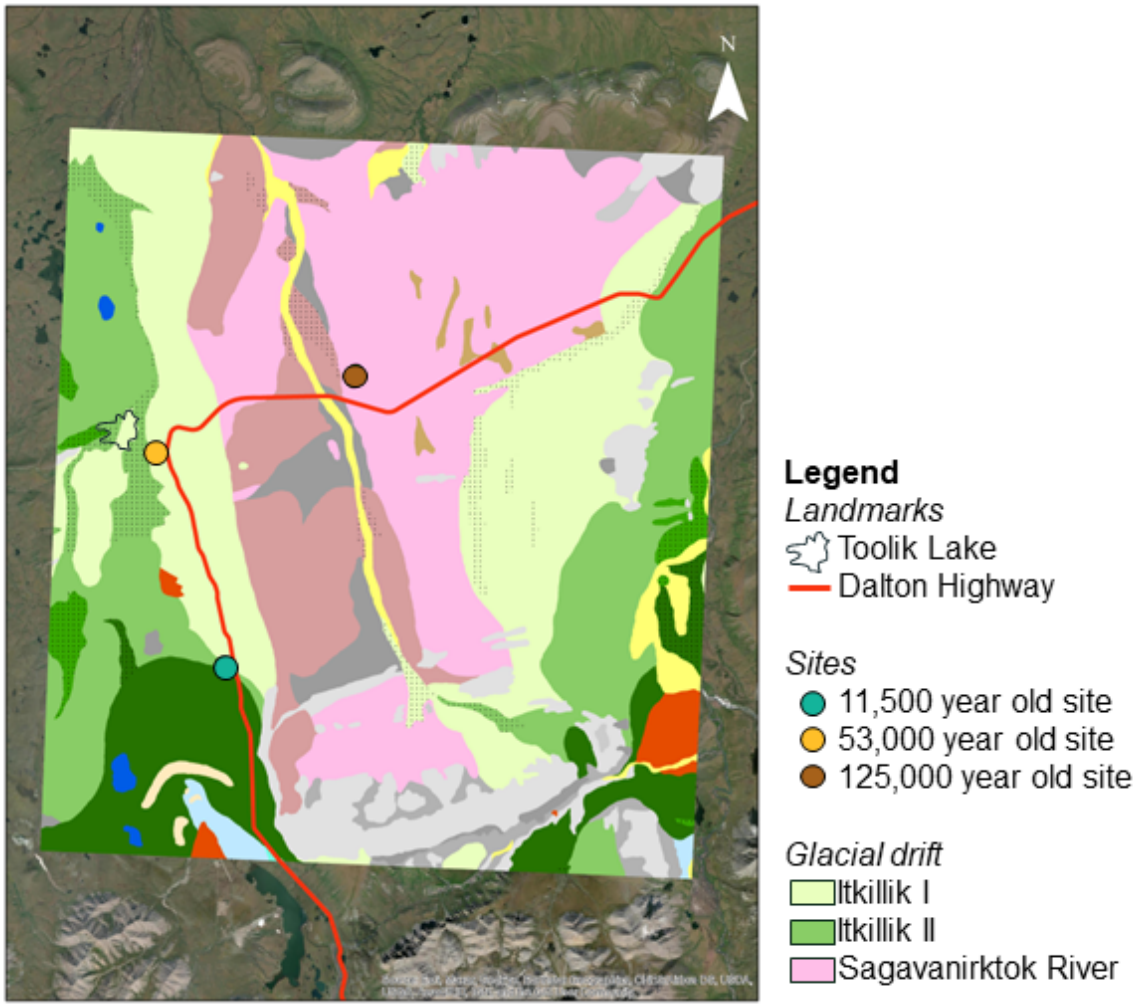
Previous work by former UNH master's student D.V. Bakke in Dr. Jessica Ernakovich's lab investigated microbial communities in Arctic soil using microbial DNA extractions but methodological limitations and sequencing failures in Bakke's original sample set prevented robust statistical conclusions. To address these sampling issues, Bakke's soils were re-sequenced in the fall of 2025 by the Ernakovich lab, generating two new DNA datasets (RunA and RunB). The primary goal of the present study is to evaluate these new datasets and determine whether data from the two sequencing runs can be reliably combined for downstream analyses. Specifically, I ask whether the distributions of samples produced by RunA and RunB are similar enough to justify combining them. The analysis completed thus far involved a laborious process of data quality assurance, standardization of sequencing depth via rarefaction, and visualization through ordination and diversity analyses.

Building on these microbial datasets, the aim is to provide a foundation for future studies exploring the influence of glacial drift and sample depth on permafrost microbial communities in Alaska. By re-assessing Bakke's samples with improved data quality and analytical rigor, this work establishes a framework for investigating the drivers of permafrost microbial diversity and informs predictions of microbial-mediated carbon flux under a warming Arctic.

Methods:

Sample collection and History

Permafrost and active layer soil samples were collected from three distinct study sites on three different glacial drift deposits near Toolik Field Station, Alaska, in August 2021 by a team from UNH. Site soil materials range in age from 11,500 years to 125,000 years. Samples evaluated in this analysis are sourced from the oldest sampling site. See map below created by Fernando Montano Lopez for sample locations and glacial drift informatio (retrieved from Bakke, 2024).



Data Pre-Processing

PCR sample products were sent to the UNH Hubbard Center for Genome Studies during the fall of 2025 and sequenced to 2x250 base pare pair-end sequencing using Illumina Novaseq 6000. Amplicon sequencing data received back from UNH Hubbard Center was fed through the Ernaklovich lab's data2 pipeline on UNH's Premise supercomputer, transforming amplicon results and assigning ASV taxonomy to individual samples using Silva db v138 (Callahan et al. (2016), Holland-Moritz et al. (2023)). Data2 pipeline outputs were then fed into R and RStudio software for all additional manipulation.

Data Processing in R and RStudio

Data were imported into RStudio for processing and analysis. After initial data cleaning, the first step assess whether any samples were present in both RunA and RunB, as overlapping samples would allow for direct statistical comparison between sequencing runs. Unfortunately, no overlapping samples were detected, meaning that RunA and RunB contain completely distinct sets of samples. This absence of sample overlap limits the comparative analyses that can be performed, as any statistical differences detected between runs cannot be disentangled from underlying sample variation. Consequently, analyses were conducted separately for each run when necessary, with visual and descriptive comparisons used to evaluate sequencing depth, diversity, and community composition.

Sequencing depth was examined for each run and the data were standardized through rarefaction to create fair comparisons of microbial richness and diversity across samples. Ordination analyses were then performed using Bray–Curtis dissimilarity metrics, visualized through Principal Coordinates Analysis (PCoA) and Non-metric Multidimensional Scaling (NMDS) to assess patterns of community similarity. PERMANOVA was also applied to test for differences in community composition between runs; however, statistical results should be interpreted with caution due to the absence of overlapping samples between RunA and RunB, meaning observed differences may reflect distinct sample sets rather than true differences in microbial community.

Results and Discussion

Examining the distribution of sequencing depth, RunA and RunB exhibited substantial differences. The majority of samples in RunA had fewer than 50,000 reads (Figure 1), whereas RunB samples generally had much higher read counts, often hundreds of thousands (Figure 2). These differences in read depth are important to consider because they influence diversity analyses and ordination results. See RunA (above) and RunB (below) data distribution summary below highlighting these differences.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8    4432   11886   21441   30466  150788
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 370  183675  299500  279413  369112  610004
```

Comparing raw sequencing counts between two different sequencing runs directly is inherently biased, as samples with higher read counts (ie. RunB) will appear to contain an increased diversity of microbial taxa. To enable fair comparisons across RunA and RunB, normalization is required; in this study, this was accomplished through rarefaction. Rarefaction involves selecting a target read depth (in this case, 1,000 reads per sample) and randomly subsampling each sample to that depth. Samples with fewer reads than the target are typically excluded,

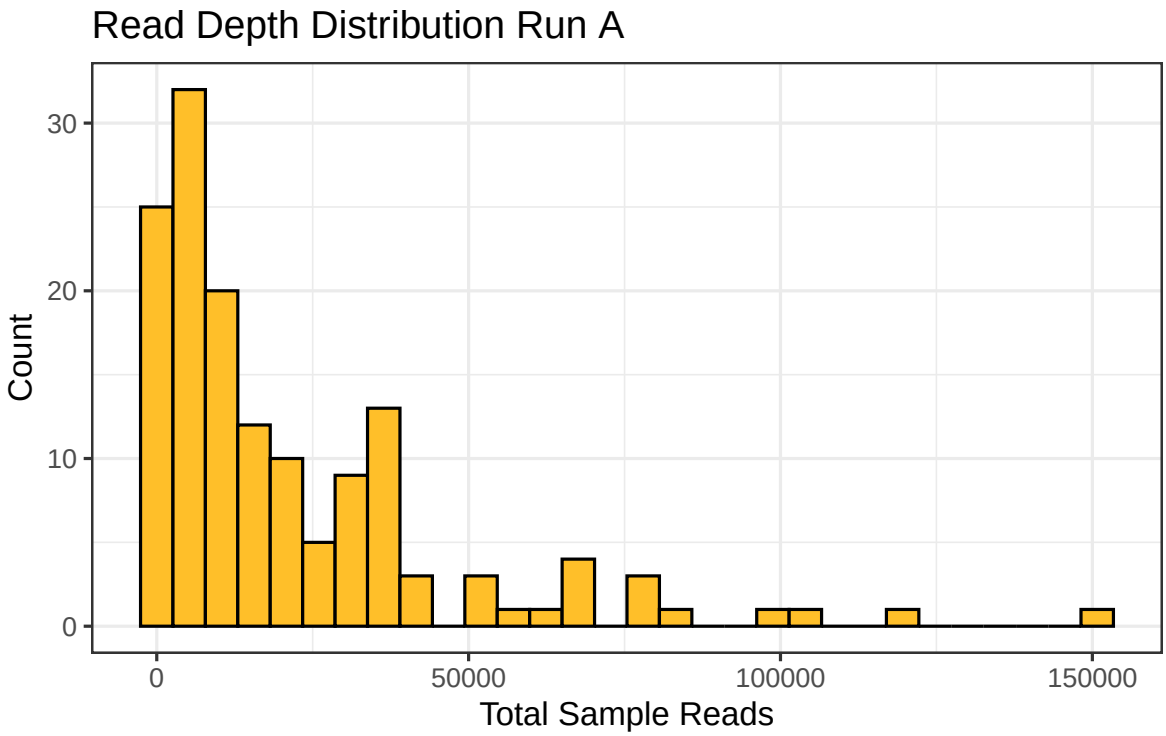


Figure 1: RunA DNA reads distribution depth, majority of samples have less than 50000 reads.

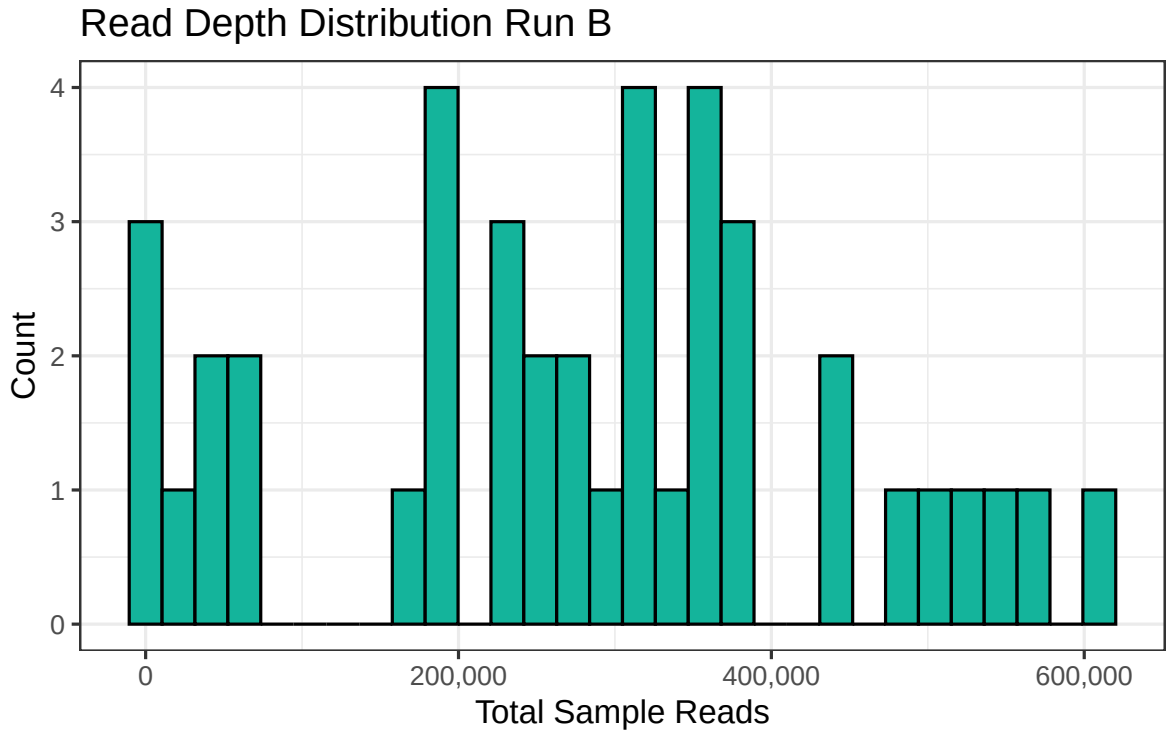


Figure 2: RunB DNA reads distribution depth, RunB is much more deeply sampled than RunA with many samples having hundreds of thousands of reads.

in this case 8.56% of samples are dropped during rarefaction. This subsampling procedure results in a dataset in which all retained samples have the same total read count, allowing unbiased comparisons of microbial richness, diversity, and community composition. Although a rarefaction depth of 1,000 reads is lower than what is typically recommended (5000 reads is often considered a minimum), it was chosen here to maximize the number of samples retained. Since the primary goal of this analysis is to determine whether the datasets from RunA and RunB can be combined, prioritizing sample retention over capturing rare taxa was appropriate for this pre-analysis. See Figure 3 for combined sample reads with rarefaction level.

To examine patterns of microbial community composition across samples, I calculated dissimilarities using the Bray–Curtis index, a metric that quantifies differences in community composition based on the relative abundance of taxa. These dissimilarity matrices were then visualized using two ordination methods: Principal Coordinates Analysis (PCoA) (Figure 4) and Non-metric Multidimensional Scaling (NMDS) (Figure 5). PCoA provides a metric-based ordination that preserves the actual distances between samples (as much as is possible in a reduced number of dimensions) while NMDS is a rank-based approach preserving the rank order of dissimilarities. Together, these ordination techniques allow for the visualization of community similarities and differences among samples, helping to identify potential clustering by sequencing run, sample depth, or other environmental factors.

Both the PCoA and NMDS analyses visually suggest substantial differences between the microbial communities represented in RunA and RunB. Unfortunately, statistical comparison of these runs is fundamentally limited because, again there are no overlapping samples between the runs. Despite this limitation, I conducted a PERMANOVA in an exploratory capacity, with the goal of determining whether the analysis could offer any additional insight into the data or help guide next steps for evaluating the two sequencing runs. PERMANOVA results below.

```
Permutation test for adonis under reduced model
Permutation: free
Number of permutations: 999

adonis2(formula = combined_bc ~ Run, data = combined_plot_df)
          Df SumOfSqs      R2      F Pr(>F)
Model      1    6.129 0.07613 15.245  0.001 ***
Residual 185   74.377 0.92387
Total    186   80.506 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Analysis of Variance Table

Response: Distances
```

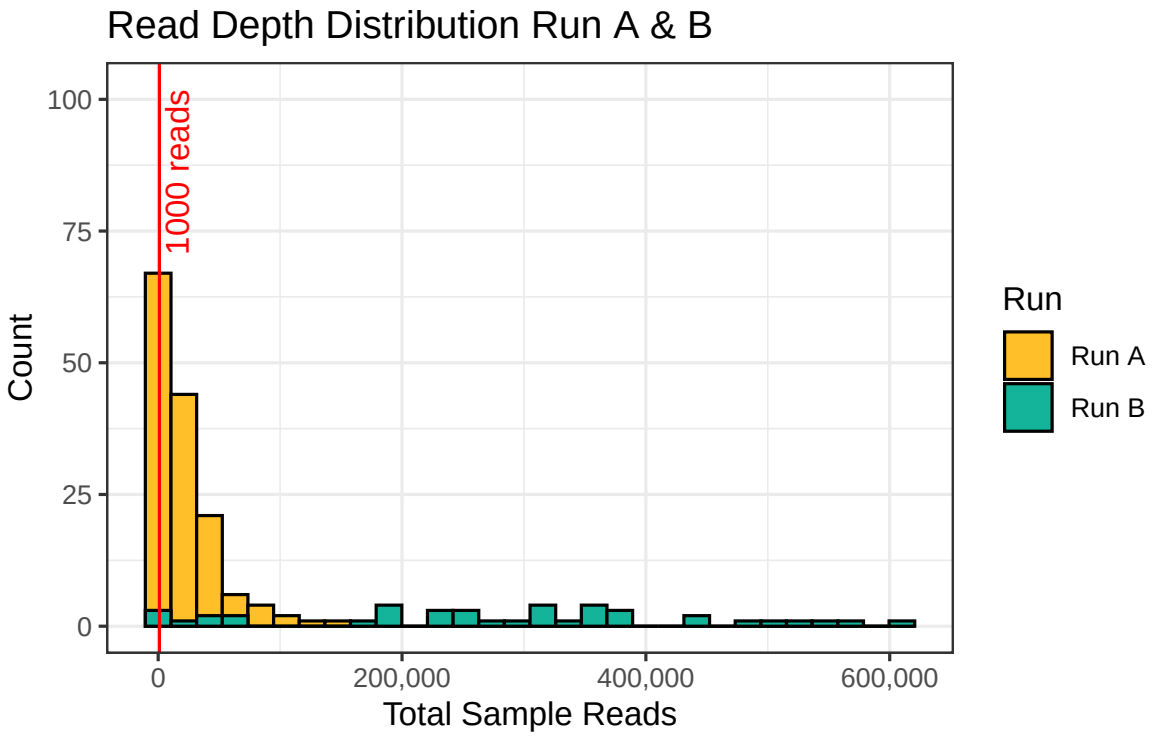


Figure 3: Combined sequencing depth with rarefaction level denoted at 1000 reads.

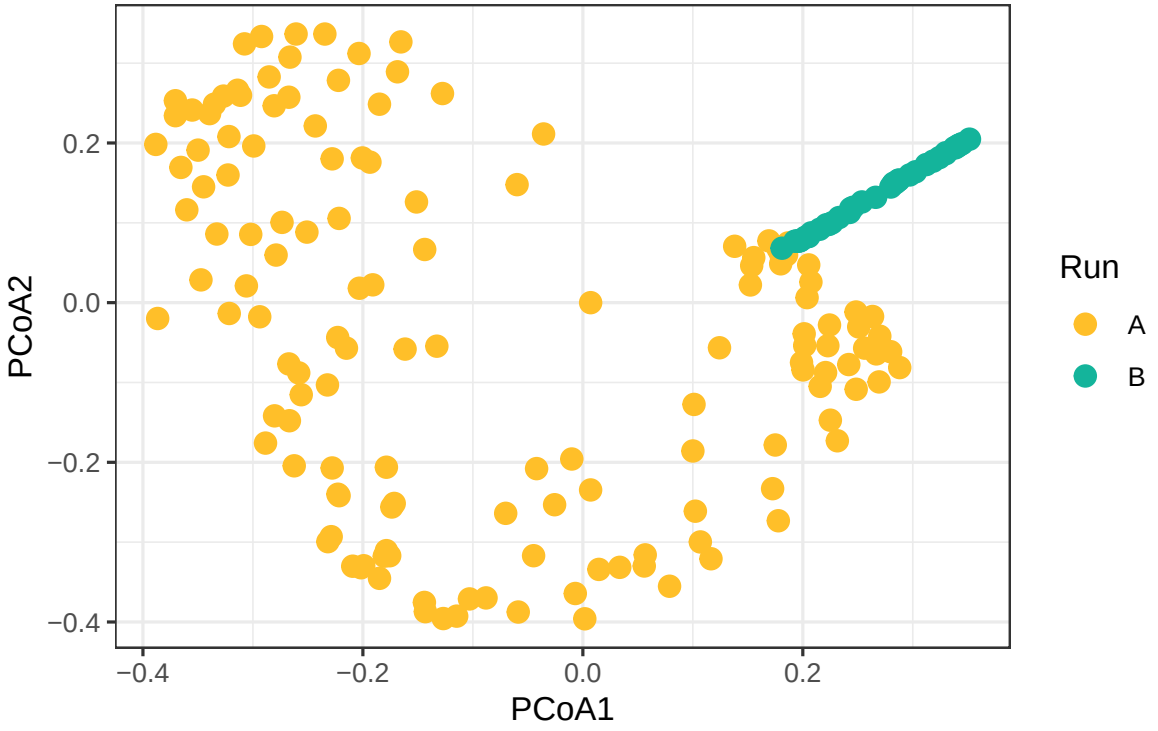


Figure 4: PCoA of Combined Samples.

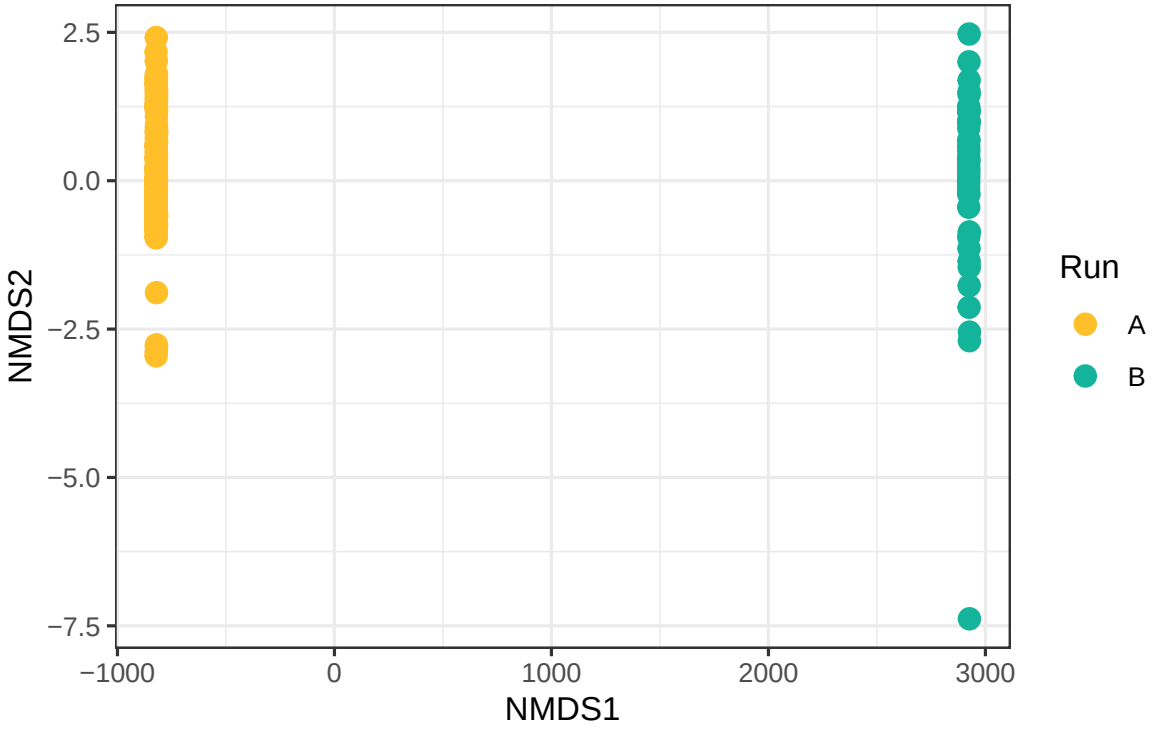


Figure 5: NMDS of Combined Samples.

```
           Df Sum Sq   Mean Sq F value Pr(>F)
Groups      1 0.0129 0.0128972  1.7825 0.1835
Residuals 185 1.3385 0.0072353
```

Results from the PERMANOVA suggest that RunA and RunB show a statistically significant but weak difference in community composition (PERMANOVA $R^2 = 0.075$, $p = 0.001$), and betadispersion confirmed that within-group variances do not differ ($p = 0.256$). However, because the two runs contain completely distinct sets of samples, this result remains inconclusive - further emphasizing that that overlapping samples are the key to determining if these two data runs can be combined.

Conclusions

The primary goal of this analysis was to evaluate two newly generated sequencing datasets (RunA and RunB) derived from the re-sequencing of soil samples originally analyzed in D.V. Bakke's master's thesis Bakke (2024). Through examination of sample read-depth distribution, rarefaction, and plotted ordinations, this work attempts to address whether the two runs could be reliably combined for downstream ecological analyses. The results suggest clear visual differences between RunA and RunB in sequencing depth and community structure; however, because the runs share no overlapping samples, it is impossible to distinguish whether these differences reflect true biological variation or variation elsewhere along the sampling pipeline. The current evidence suggests that the two datasets should not be combined.

Moving forward, the most critical next step is generating overlapping samples across sequencing runs so that said variation can be statiscally evaluated. Establishing sequencing overlap will determine whether RunA and RunB can ultimately be merged into a unified dataset suitable for re-analyzing Bakke's original ecological hypotheses related to glacial drift and soil depth. In the meantime, each run can be futher analyzed independently to investigate within-run ecological patterns, including Shannon and Simpson diversity, which can provide additional insight into microbial community structure and variability within each dataset.

Bakke, Dana Victoria. 2024. "Characterizing the Influence of Depth and Glacial Drift on Microbial Community Diversity and Potential Carbon-Cycling Enzyme Activity in Permafrost Soils."

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Holland-Moritz, h, A Oliverio, c Walsh, m Gebert, and k fan. 2023. *Dada2 Tutorial with NovaSeq Dataset for Ernakovich Lab. Github Repository.* Github Repository. https://github.com/ErnakovichLab/dada2_ernakovichlab.

Mack, Michelle C., Edward A. G. Schuur, M. Syndonia Bret-Harte, Gaius R. Shaver, and F. Stuart Chapin. 2004. "Ecosystem Carbon Storage in Arctic Tundra Reduced by Long-Term Nutrient Fertilization." *Nature* 431 (7007): 440–43. https://doi.org/10.1038/nature02887.

Rantanen, Mika, Alexey Yu Karpechko, Antti Lipponen, Kalle Nordling, Otto Hyvärinen, Kimmo Ruosteenoja, Timo Vihma, and Ari Laaksonen. 2022. "The Arctic Has Warmed Nearly Four Times Faster Than the Globe Since 1979." *Communications Earth & Environment* 3 (1): 168. https://doi.org/10.1038/s43247-022-00498-3.

Schuur, E. a. G., A. D. McGuire, C. Schädel, G. Grosse, J. W. Harden, D. J. Hayes, G. Hugelius, et al. 2015. "Climate Change and the Permafrost Carbon Feedback." *Nature* 520 (7546): 171–79. https://doi.org/10.1038/nature14338.

Schuur, Edward A. G., Benjamin W. Abbott, Roisin Commane, Jessica Ernakovich, Eugenie Euskirchen, Gustaf Hugelius, Guido Grosse, et al. 2022. "Permafrost and Climate Change: Carbon Cycle Feedbacks From the Warming Arctic." *Annual Review of Environment and Resources* 47 (Volume 47, 2022): 343–71. https://doi.org/10.1146/annurev-environ-012220-011847.