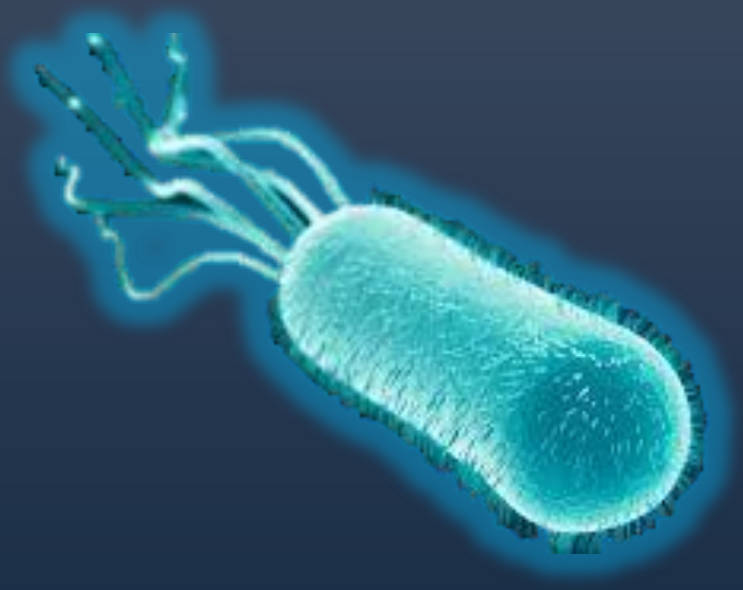
2024



# A Microbiological Introspective into River Biota Composition and Diversity Utilising Amplicon Multivariate Analyses

CHRISTIAN PEPLER - 24065072

**WORD COUNT: 2,496 WORDS (EXCLUDING ABSTRACT)**
ABSTRACT WORD COUNT: 206 WORDS

## Abstract

Microbial communities are core factors in carbon regulations within river ecosystems. Such metabolic activity is influenced by river depth, flow and conditions surrounding the environment. This is typically predicated on dissolved organic carbon (DOC) content causing damage to water treatment facilities, incurring extremely high costs to local councils. Using simulated river channels for peat and sugar exposure alongside bacterial swabbing with situated ceramic tiles and existing stones, amplicon sequencing can elucidate bacterial composition, dispersion metrics and relationships in diversity that may underpin core behavioural activity of river microbiota. Through pre-processing and statistical modelling of NMDS plots, heat trees and a randomForest, the present aims were to outline diversity differences in tile collections and dispersions in treatment conditions. Results indicated high significance but poor reliability and $R^2$ value explanations, yet figure inferences portrayed large compositional dominance on the bottom of the tile and stone pickup. Dispersions skewed axially for sugar and river treatment variables, expressing subtle but insignificant changes in bacterial presence. Biologically, bacteria favoured the bottom of the tile, potentially due to dependency on hypoxic zones for rich metabolites. This may indicate spatial preferences compared to metabolic sources. Frequent disparities and high clustering are observed, hence analytical limitations are discussed and solutions and replications are suggested.

## Contextual Background

Microbiome communities residing in river ecosystems are often impacted by water quality due to dependencies upon the cycling of chemical compounds like nitrogen, sulfur, carbon and phosphorus. Yet, microbiota can act as amplifying organisms to the very same water quality – often observed mediating metabolic activity like environmental processing and xenobiotic degradation (Breton-Deval, 2020). This is why some bacterial communities present almost symbiotic relationships with their environment. Such relationships are distinctly varied from each other in aerobic vs anaerobic microbes that regulate organic matter in differing oxygen-dense zones. These zones are impacted by environmental factors such as depth, flow and decomposition rates, allowing for species diversity of microbiota communities to increase or decrease dependent on favourable conditions (Laudon *et al*, 2011).

Carbon is one of the core compounds regulated within river ecosystems. Integral to life in general, its cycle remains prevalent in these communities. Particularly in surface oxygen-rich regions, dissolved organic carbon (DOC) is extensively metabolised for energy in respiration and production of increased biomass. As such, anaerobic counterparts catalyse DOC to

produce a greater oxygen depletion for resultant methane and various reduced compounds. Great sources of DOC stem from Peat and Sucrose, the prior highlighting a slower release of compounds due to time-consuming processes, the latter being common for microbial fermentation due to its disaccharide carbon-dense molecular structure. Proposing these two conditions and measures of composition to a simulated river can increase our understanding of microbial behaviour and preventing harmful conditions in environmental anthropogenic activities.

Due to the present regulatory factors of DOC content upon microbiota community thriving, the potential for conditional differentiation increases. This can allow for distinguishment of advantageous bacteria present in these simulations, implicating manual regulation of water quality, especially on drinking water reserves, agriculture and peatland drainage. These interventions may continually prevent episodic shutdown of water treatment facilities. Furthermore, avoidance of cross-stream concentrations is integral to discerning microbiota activity in DOC lysis.

A simulative and metagenomic 16S rRNA analysis will occur on multiple control, peat-exposed and sucrose-exposed (randomly assigned between sites as to not favour lateral sedimentation) bacterial communities in differing upland streams (L3, L6, L7) where 16 artificial channels replicate conditions using re-circulation technology to novelly investigate isolated differences. The objectives of the project are as follows:

- Determine microbiota differences between tile orientation – (hypoxic (depth - bottom) vs hyperoxic (surface - top) zones).
- Identify differences or similarities between treatment conditions.
- Outline differences between stream diversity.

The alternate hypotheses are:

- **H1** = The tile recording biota community in hyperoxic zone will detect a higher diversity than its hypoxic counterpart (more aerobic bacteria present).
- **H2** = The microbiota composition will be more diverse in sugar exposure.
- **H3** = There will be a significant difference between stream diversity L3, L6 and L7.

## Methods

For the purpose of this study, amplicon sequencing analytical methodology is of paramount importance. Data was sequenced using a PCR assay after separation from tile

sedimentation. Microbiota 16S rRNA sequences remained singularly forward reads throughout the process.

**Pre-Processing**

Microbiota data was collected from and compiled into 97 sets. Utilising *Linux*, MobaXTerm and the programming language BASH, data was trimmed to prevent noise and run a multiqc test using fastp to detect and visualise the quality of the data (Ewels *et al*, 2016). This ensures that reads with high confidence are maintained for a downstream analysis method to fundamentally improve the phred (quality scores = QC) scores and filter PCR-amplified data. The generated multiqc report allows for early detection of issues in the data preventing skewness and false outputs.

A manifest file was employed to label the metadata and match up sequences with the 97 samples for cross-sectional integration. Using a tab delimited format (tsv) separated datum between columns can aid reading proficiency within the subsequent bioinformatical steps. A “single-end-demux.qza” file was outputted and utilised with a script to merge IDs to samples. The qiime2/ module was loaded into the Linux console for denoising (Bolyen *et al*, 2019). The trimmed fastq.gz files outputted from the trimming process were input into the DADA2 plugin for resultant further trimming of noise in the high through-put data (Callahan *et al*, 2016). No further truncation was undertaken as a result of high quality but large at-end oscillations in scores were observed in the multiqc report, prompting further trimming of 8 forward read sequences. Further ambiguity in sequence returns were removed, including duplications, chimeras and artifacts. Generated table and replicated sequences qza files that contained ASVs (Amplicon Sequence Variants - standardised single nucleotide differentiation format) supported taxonomic analysis within the qiime2/ module, with more clarity as compared to the OTU (Operational Taxonomic Unit) format clustering (Chiarello *et al*, 2022).

Taxonomic classification adopted the use of a SILVA classifier within a machine learning tool. With the classify-sklearn command and the relevant taxonomy information (“silva-138-99-nb-classifier.qza”), the replicate sequences are mapped against the labels to produce a “taxonomy.qza” file. This is required to fulfil a metadata tabulate command producing qiime2/ compatible visualisation (qzv file). Further inclusion of the raw stream metadata emits a taxa bar plot, integral to predicting biodiversity metrics and relation of communities to environmental factors – of which are dependent variables. Bacterial identification gives genus, phylum and class information, recording metabolic involvement and top microbiological contribution from each presence.

**Statistical Data Analysis**

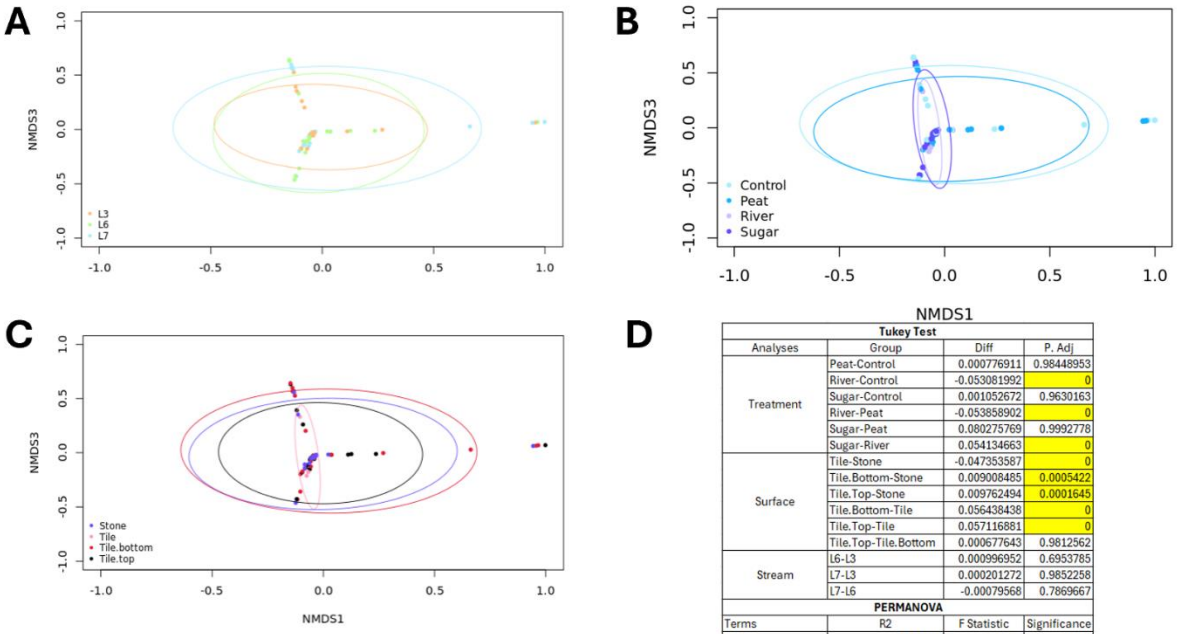
Further statistical and model construction occurred in the R Studio (R Core Team, 2024) online version bridging all required outputs from MobaXTerm through the intermediate of posit.sponsa. Several installations were required for vibrant visualisation and effective operational analysis. The following were employed: *devtools* – for its compatibility with GitHub and debugging packages to streamline workflow (Wickham *et al*, 2022); *metacoder* – paramount for metagenomic analysis and visualisation of microbial community composition (Foster *et al*, 2017); *vegan* – comprehensive package for ecology containing diversity indices tools and powerful statistical analysis commands (adonis2) (Oksanen *et al*, 2024); *randomForest* – contains an algorithm specializing in regressions, predictions and highly appropriate for machine learning models (Breiman *et al*, 2024); *qiime2R* - to import qiime2 output for compatibility (Jordan, 2018). The metadata is read in and a colour palette set alongside ASV containing feature tables. Final metadata formatting checks took place.

NMDS (Non-metric Multidimensional Scaling) ordination technique was employed to benefit a Bray-Curtis model (or an emperor plot) in assessing dimensionality of microbial community data. A rarefaction curve was generated for quality checking of species richness in combination with sufficient sequencing – observed within the plateau of the graph. Rarefaction is commonly used to standardise ranging diversity species in ecological data with vast outputs. It achieves uniformity and minimises bias by discarding data which potentially becomes necessary in high through-put methods and complex data sets like the present samples. Data was resultantly rarified to 2707 reads and transformed (fourth root) to control abundance in sequence depth. Heat trees were constructed to present microbial community composition pertaining to taxa abundance, hierarchy and presence between conditions in an aesthetic, communicative format. Random forest plot was also constructed to show an element of prediction with the surface/tile orientation data.

Statistical analysis operated around a PERMANOVA (Permutational Multivariate Analysis of Variance). Management of continuous and categorical variables are necessary for hypothesis testing due to the range of data observed and type of data being examined. Computational efficacy is attributed to the utilisation of a random element for estimation modelling. The statistical assumption remains that the dispersion is not significantly different between groups/conditions. A supportive Tukey test was completed to identify significant differences between variables, testing the assumption. Furthermore, p-value adjustments are made to prevent false positives (Type 1 errors). The alpha numeric was set to align with a calculated metadata gradient to manage within-sample diversity across differing read counts in the collection.

## Results

Results are formulated into two multi-panel figures (*Figure 1* and *Figure 2*) and conventionally reported statistical output. Data was sufficiently explored but main focus remained upon Surface variables.




**Tukey Test**

| Analyses | Group | Diff | P. Adj |
|---|---|---|---|
| Treatment | Peat-Control | 0.000776911 | 0.98448953 |
| | River-Control | -0.053081992 | 0 |
| | Sugar-Control | 0.001052672 | 0.9630163 |
| | River-Peat | -0.053858902 | 0 |
| | Sugar-Peat | 0.080275769 | 0.9992778 |
| | Sugar-River | 0.054134663 | 0 |
| Surface | Tile-Stone | -0.047353587 | 0 |
| | Tile.Bottom-Stone | 0.009008485 | 0.0005422 |
| | Tile.Top-Stone | 0.009762494 | 0.0001645 |
| | Tile.Bottom-Tile | 0.056438438 | 0 |
| | Tile.Top-Tile | 0.057116881 | 0 |
| | Tile.Top-Tile.Bottom | 0.000677643 | 0.9812562 |
| Stream | L6-L3 | 0.000996952 | 0.6953785 |
| | L7-L3 | 0.000201272 | 0.9852258 |
| | L7-L6 | -0.00079568 | 0.7869667 |

**PERMANOVA**

| Terms | R2 | F Statistic | Significance |
|---|---|---|---|
| Surface | 0.01987 | 0.9242 | 0.998 |
| Stream | 0.02213 | 1.0292 | 0.165 |
| Treatment | 0.224 | 1.0421 | 0.071 |
| Stream:Treatment | 0.04479 | 1.0407 | 0.033 |
| Surface:Treatment | 0.04303 | 0.9998 | 0.494 |
| Surface:Stream | 0.0408 | 0.948 | 0.999 |

*Figure 1* - Multi-Dimensional Scaling and Statistics. **A:** NMDS Bray-Curtis plot presenting dispersion between Streams: L3 (Orange), L6 (Bright Green) and L7 (Light Blue) with respective ellipses identifying clustering. **B:** NMDS Bray-Curtis plot presenting dispersion between Treatments: Control (Light Blue), Peat (Dark Blue), River Collection (Lilac) and Sugar (Purple). **C:** NMDS Bray-Curtis plot presenting dispersion between Surface collections: Stone (Purple), Tile (Pink), Tile.Bottom (Red) and Tile.Top (Black). **D**: Table presenting statistical output for Tukey assumption test (Group, Difference score and adjusted p value) and PERMANOVA ($R^2$, F Statistic and Significance score: $p < 0.05$ = Yellow, $p < 1.0$ = Gold) presenting degree of influence upon the dataset with interaction variables listed. F Statistics are reported to 3 significant figures in text.

The Tukey assumption test manifested high significance in Treatment and Surface groups. Treatment groups River-Control ($p < 0.0001$), River-Peat ($p < 0.0001$) and Sugar-River ($p < 0.0001$) revealed negative assumption congregation, suggesting substantial differences in bacterial composition. Surface groups Tile-Stone ($p < 0.0001$), Tile.Bottom-Stone ($p = 0.0005422$, $p < 0.001$), Tile.Top-Stone ($p = 0.0001645$, $p < 0.001$), Tile.Bottom-Tile ($p < 0.0001$) and Tile.Top-Tile ($p < 0.0001$) indicate extreme variance between the sourced stone from the river and also the prematurely placed tile, of which top and bottom treatments are highly distinguished. All other results met non-significant assumptions for the PERMANOVA.

The PERMANOVA displayed significance in the individual term Treatment ($F_{1.042}$, $R^2$ = 0.224, p = 0.071, p < 1.0) and the interaction term Stream: Treatment ($F_{1.041}$, $R^2$ = 0.04479, p = 0.033, p < 0.05). These suggest significance in Treatment conditions upon microbiota in each stream due to significance popularity with treatment. Yet, inconclusive $R^2$ values depict an extremely low explanation for the relationship of treatment upon other variables. This further indicates high clustering and similarity of microbiota composition within the simulations (*Figure 2A*), confounding relationships observed between ceramic tile collection and stone-river collections.

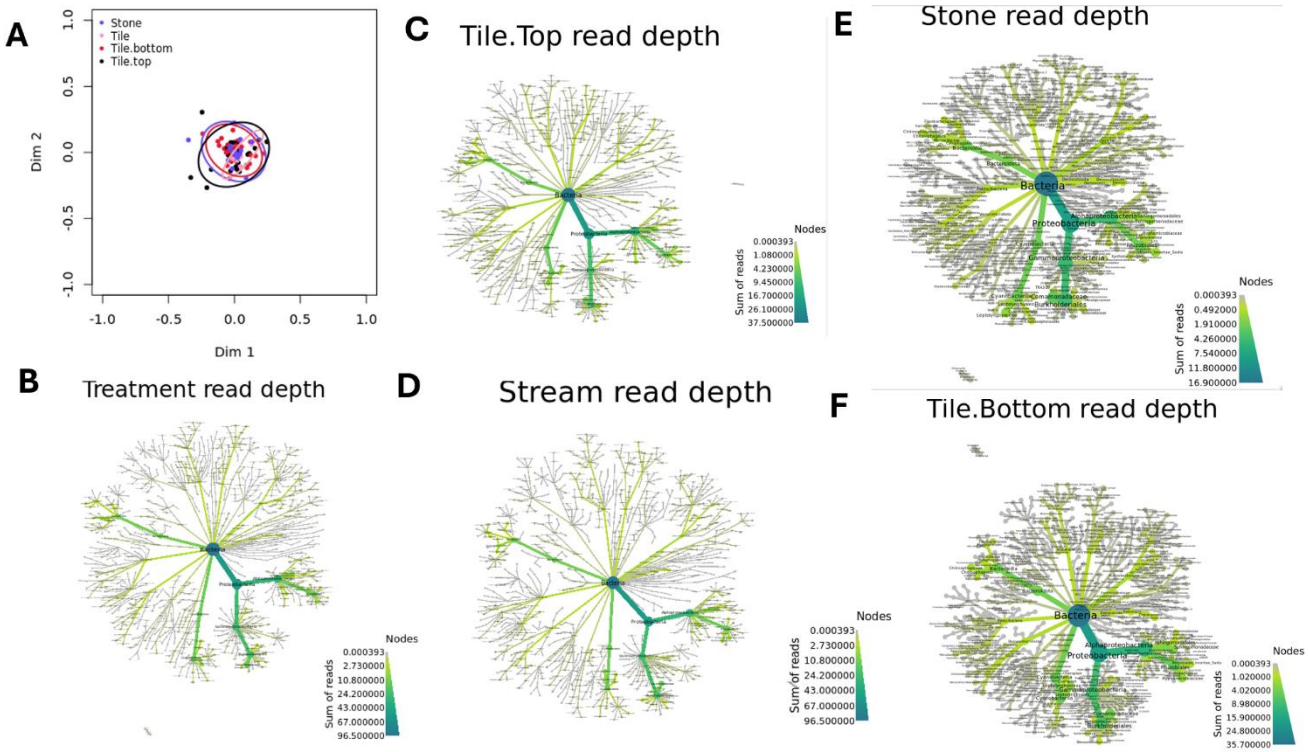



*Figure 2* - Composition and Clustering. **A**: Random Forest machine learning predictor plot representing effectiveness in assigning variables to each Surface group based on data with respective ellipses presenting clustering distance. Colour code presents Stone (Purple), Tile (Pink), Tile.Bottom (Red) and Tile.Top (Black). **B**: Heat Tree outlining microbial composition of Treatment bacterial sequences where thickness and darker colour portray increased bacterial prevalence. **C**: Heat Tree outlining microbial composition and clade differentiation of Tile.Top (hyperoxic) bacterial sequences where thickness and darker colour portray increased bacterial prevalence. **D**: Heat Tree outlining microbial composition of Stream bacterial sequences where thickness and darker colour portray increased bacterial prevalence. **E**: Heat Tree outlining microbial composition and abundance for solely Stone collection bacterial sequences where thickness and darker colour portray prevalence levels. **F**: Heat Tree outlining microbial composition and abundance for solely Tile.Bottom (hypoxic) collection bacterial sequences where thickness and darker colour portray prevalence levels.

Each heat tree showcases high compositional similarity, par Surface collections, potentially discerning that varying conditions have little effect upon community variation. Furthermore, high clustering and error rates resulting from the Random Forest plot (*Figure 2A*) enforce this phenomenon. Proteobacteria, containing alpha and gamma classes, are highly represented in all trees, especially dense in Tile.Bottom and Stone (*Figure 2B, 2C, 2D, 2E and 2F*).

## Discussion

### Surface Collection Dynamics

What becomes surprising is the observations within Figure 2A with high clustering and no separation of Tile collection. Error rates are obviously high from the machine learning model, indicating poor ability to separate data into correct groups, despite Tile previously being distinct from other clustering patterns (*Figure 1C*). This compositionally suggests poor distinguishment between collections, other than a pairwise Tukey analysis of Stone collections against Tile.Top and Tile.Bottom. Interpretations elucidate that the stone withholds a much larger set of microbial diversity and greater abundance of reads than its ceramic counterparts, as proven by Figure 2E. With greater clade density, it becomes more alarming with a lack of reliable significance from Surface PERMANOVA interactions. Logically, as the stone has remained in the river for an unrecorded amount of time, bacterial respite has embedded itself upon it.

In relation to hypoxic and hyperoxic zones, it appears that composition and abundance is much higher in the hypoxic zone indicated by Figure 2F. Furthermore, observed differences in Figure 1D between Stone collection and Tile.Bottom represent very small variations from the two. It is with this data that H1 can be rejected for its opposing outcome. Increasingly, *Burkholderia* represents a high taxonomic revelation which becomes concordant with the taxa bar-plot generated in pre-processing. Interestingly, large portions of this genus are aerobically regulated and contain pathogenic properties (Sass *et al*, 2013). Yet, it is possible that bacteria favouring anaerobic metabolic cycles are fostered with this finding, where a potential dependency forms upon the acid rich soil and carbon treatments for ATP generation (Curiel Yuste *et al*, 2007). It is here that DOC metabolism can be fostered, providing energy and sustainability for hypoxic bacterial communities. Through this feedback system, community construction is proposed to occur more readily within deeper, spatially varied streams as represented with the comparison between Figure 2C and 2F (Nogaro *et al*, 2013). Regardless, debating factors are highlighted as the disparity between Figure 2A and its compositionally clear, data-equivalent peers Figure 2C, 2E and 2F are unavoidable.

**Microbiota Between Treatments and Streams**

Upon interpretation, between treatment differences are highly alike. This is reflected in the degree of dispersion (*Figure 1B*). The ellipses are closely contained indicating low variance, except for the Sugar and River plot which differentiates upon a separate axis. Furthermore, this low variance continues in statistical endeavours particularly with Treatment in the PERMANOVA (*Figure 1D*). Biological inferences manifest extremely similar composition in microbiota, subtle change is observed but not to a notable degree. Increasingly, expressions to this degree are mirrored in Figure 2C, 2B and 2D heat trees where Proteobacteria remains extremely dominant, yet low node areas differ sporadically. Bacterial architecture within these communities seem to align both within treatment conditions and streams (*Figure 2B and 2D*). It is with this observation that H3 can be rejected.

Tukey assumption tests contextually indicated high significance between all River interactions. This may be due to the brief collection of the “Tile” (designated “River” in the metadata) in each stream (L3, L6 and L7) through a Day and Night period. This difference between collection sample can explain the NMDS plots in Figure 1B and 1C simultaneously, as patterned by the coordinated ellipses in each graph. Biologically, a lower diversity in the microbiome would be statistically observed. However, repeated high significance is reported. This queries the biota present between groups, thus leading to a rejection of H2. Sugar exposure impacts the biota dynamics to a minimal degree, but methodology could be a convoluted factor here. It is possible that treatment differences and stream orientation do not effect bacterial dependencies to the same degree as river localisation (hyperoxic vs hypoxic zones). Thus, DOC content manipulations could fall short of DOC source location manipulations.

It is incredibly apparent to not overstate the importance of these findings. Drawbacks require sufficient attention alongside their limiting qualities.

**Critical Reflection**

Limiting factors dictate a seemingly significant effect upon data extrapolation. Specifically the randomForest machine learning model has been outlined to be very susceptible to overfitting and poor validation, especially within close to non-linear relationships (Hengl *et al*, 2018). Furthermore, PERMANOVA influences significance score due to a random element within the computational approach – therein effecting true statistical outcome. Upon reflection towards analytical integrity, exclusion of particular datums could have manifested differing results such as the “Tile” Day and Night collections for significance revelations pertaining to other Surface parameters. Other limiting factors include inaccuracies to specific trimming measures when controlling quality scores in pre-processing, lack of reliability due to no post-

hoc test (Tukey tested assumption parameters) and difficulty in outlining species level variation from heat tree generation.

**Contextualisation**

In relation to concurrent research, this data produces both positive and negative connotations. Hugerth *et al*. (2017) describes an overwhelming fondness for identification of driving forces influencing species diversity in amplicon sequencing. As such, conclusions could not be ascertained in the present study, assertive interpretations can elicit some direction. Community network interactions are also presented to be a positive outcome. However, Hugerth *et al* (2017) does focus whole-heartedly upon the adoption of OTUs, an aforementioned, lack-luster counterpart when dealing with taxonomic stratifications (Chiarello *et al*, 2022). As this article reviews amplicon sequence techniques, the authors deservingly establish artificial neural networks (ANNs) as a formidable alternative to the randomForest algorithm – providing sufficient solution to present credible drawbacks (Larsen *et al*, 2012).

## Conclusion

No concrete explanations for the lack of significance between treatments and surface collections can be made due to statistical support. Implications of this study remain minimal due to factorial unreliability and expressed possibilities of causation rather than evidenced fact. However, understanding of technique-based limitations, ecological variability in microbial river systems and bacterial community spatial behaviour which elucidates some promising replications with more streamlined outcomes.

## References

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C.,

Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R. and Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37, pp.852–857. Available at: https://doi.org/10.1038/s41587-019-0209-9.

Breiman, L. [Fortran original], Cutler, A. [Fortran original], Liaw, A. [R port, cre] and Wiener, M. [R port], 2024. Random forests. [Computer program]. Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/.

Breton-Deval, L., Sanchez-Reyes, A., Sanchez-Flores, A., Juárez, K., Salinas-Peralta, I. and Mussali-Galante, P., 2020. Functional analysis of a polluted river microbiome reveals a metabolic potential for bioremediation. *Microorganisms*, 8(4), p.554. Available at: https://doi.org/10.3390/microorganisms8040554.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, pp.581–583. Available at: https://doi.org/10.1038/nmeth.3869.

Chiarello, M., McCauley, M., Villéger, S. and Jackson, C.R., 2022. Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *PLoS ONE*, 17(2), p.e0264443. Available at: https://doi.org/10.1371/journal.pone.0264443.

Curiel Yuste, J., Baldocchi, D.D., Gershenson, A., Goldstein, A., Misson, L. and Wong, S., 2007. Microbial soil respiration and its dependency on carbon inputs, soil temperature and moisture. *Global Change Biology*, *13*(9), pp.2018-2035. https://doi.org/10.1111/j.1365-2486.2007.01415.x

Ewels, P., Magnusson, M., Lundin, S. and Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), pp.3047–3048. Available at: https://doi.org/10.1093/bioinformatics/btw354.

Foster, Z.S.L., Sharpton, T.J. and Grünwald, N.J., 2017. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLOS Computational Biology*, 13(2), p.e1005404. Available at: https://doi.org/10.1371/journal.pcbi.1005404.

Free Software Foundation, 2018. Bash manual page. *Linux*. Available at: http://man7.org/linux/man-pages/man1/bash.1.html.

Hadley, W., Hester, J., Chang, W., Bryan, J. and RStudio, 2022. Devtools: R package development tools. Version 2.4.5. Available at: https://devtools.r-lib.org/.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B. and Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, p.e5518. Available at: https://doi.org/10.7717/peerj.5518.

Hugerth, L.W. and Andersson, A.F., 2017. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Frontiers in Microbiology*, 8, p.1561. Available at: https://doi.org/10.3389/fmicb.2017.01561.

Jordan, E.B., 2018. qiime2R: Importing QIIME2 artifacts and associated data into R sessions. Available at: https://github.com/jbisanz/qiime2R.

Larsen, P.E., Field, D. and Gilbert, J.A., 2012. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, 9, pp.621–625. Available at: https://doi.org/10.1038/nmeth.1975.

Laudon, H., Berggren, M., Ågren, A., Buffam, I., Bishop, K., Grabs, T., Jansson, M. and Köhler, S., 2011. Patterns and dynamics of dissolved organic carbon (DOC) in boreal streams: The role of processes, connectivity, and scaling. *Ecosystems*, *14*, pp.880-893. https://doi.org/10.1007/s10021-011-9452-8

Nogaro, G., Datry, T., Mermillod-Blondin, F., Foulquier, A. and Montuelle, B., 2013. Influence of hyporheic zone characteristics on the structure and activity of microbial assemblages. *Freshwater Biology*, *58*(12), pp.2567-2583. https://doi.org/10.1111/fwb.12233

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.H., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C.J.F., Weedon, J., 2024. Vegan: Community Ecology Package. Version 2.6-8. Available at: https://github.com/vegandevs/vegan.

R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sass, A.M., Schmerk, C., Agnoli, K., Norville, P.J., Eberl, L., Valvano, M.A. and Mahenthiralingam, E., 2013. The unexpected discovery of a novel low-oxygen-activated locus for the anoxic persistence of *Burkholderia cenocepacia*. *The ISME Journal*, 7(8), pp.1568–1581. Available at: https://doi.org/10.1038/ismej.2013.36.