Assessing the Impact of Nuclear Contamination on Soil Microbial Communities

Heather MacGregor
Graduate Student, University of California, Berkeley
Isis Fukai[1], Kurt Ash[1], Terry Hazen[1], and Adam Arkin[2]
[1]The University of Tennessee [2]University of California, Berkeley

Introduction

- Nuclear processing activities ➜ dissemination of materials and wastes ➜ changes in physicochemical conditions
- Microbes are very sensitive to changes in their environment.[1]
  ↪ Changes in the microbial community structure and function can provide information about the physicochemical condition of its environment (Fig. 1).[2]

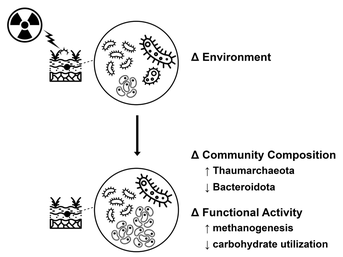

Fig. 1: Adapted from Weedon et al. (2017)

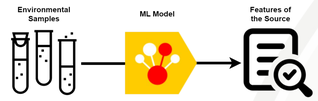

Fig. 2: ML pipeline with CatBoost.

↪ Therefore, we can infer some information about the source and characteristics of contamination in the environment by analyzing the structure and function of microbial communities. ML techniques (Fig. 2) can be applied to ID patterns and relationships that enable prediction of the type, age, and distance of the source.

Materials and Methods

Fig. 3: Map of locations associated with the 16S soil datasets used in the analysis.

- Soil sample data (Fig. 3) was sourced from MTV collaborators, publicly available datasets on the European Nucleotide Archive, and the Earth Microbiome Project (EMP).[3,4]
  ↪The EMP dataset contains 2,000+ amplicon datasets from soil samples across the globe.[4]

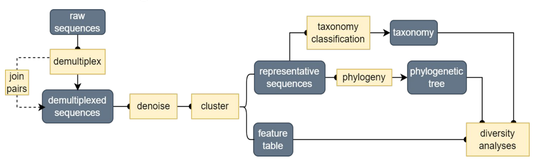

Fig. 4: Simplified QIIME 2 workflow.

- The EMP dataset is already pre-processed by QIIME, so all raw datasets were pre-processed using the same methods in QIIME 2 before analysis (Fig. 4).[5]

Mission Relevance

- These models, in conjunction with an environmental surveillance protocol, can enable remote diagnostics to be performed at locations where such activities are taking place.
- Understanding the mechanistic basis of microbial responses could aid both in detection and remediation of the signal.

Results and Discussion

- Samples cluster by dataset (Fig. 5).
  ↪ Fukushima (NPP) is more similar to ORFRC (legacy site) than Rabbit Lake (U mine)
  ↪ Subsets of SRS (legacy site) data are closest to ORFRC outliers, Cd/As- contaminated sites, and Rabbit Lake, respectively
  ↪ Lack of metadata for SRS provides a roadblock.
- Contaminated samples *appear* to cluster by pH…
  ↪ More acidic datasets (ORFRC and Fukushima) are more similar to each other than to the more alkaline dataset (Rabbit Lake).
  ↪ However, each dataset and specific site have limited pH range.
  ↪ Fukushima and ORFRC are back-end sites, while Rabbit Lake is a U mine.

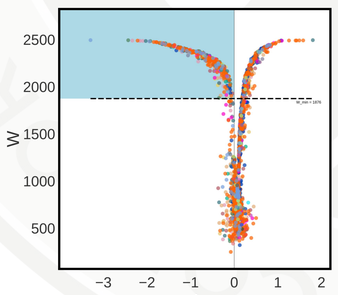

Fig. 6: ANCOM

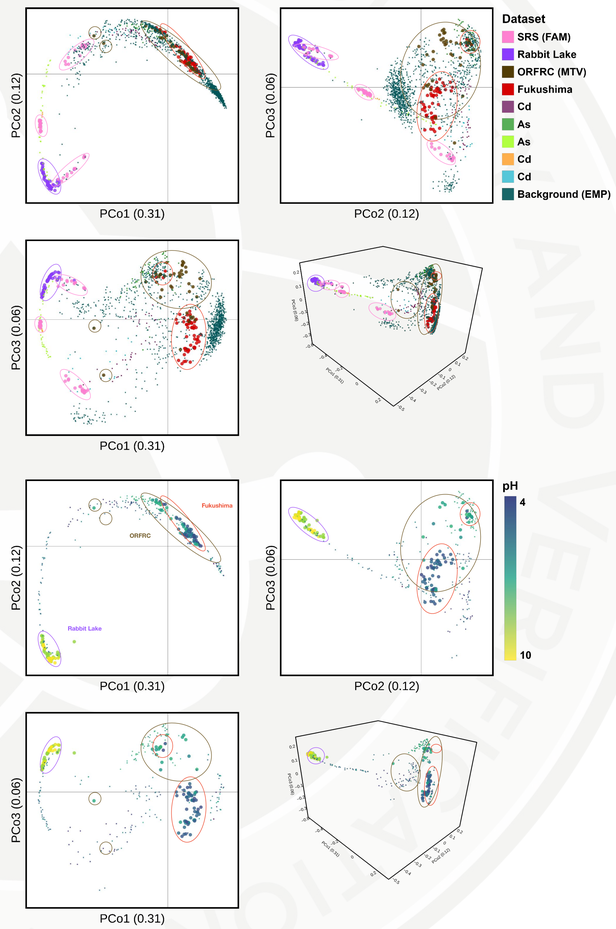

Fig. 5: Principle Coordinate Analysis (PCoA) with Bray-Curtis as a beta-diversity metric.

- At the genus level, analysis of composition of microbiomes (ANCOM) analysis identifies genera that are more abundant in contaminated or pristine samples (Fig. 6).

MTV Impact

- Integration with a community studying diverse aspects of nuclear processing and technology gives context and constraints on our work.
- The opportunity to collaborate across the university and national labs allows us access to sites, nuclear process information, and intellectual collaboration we might not have otherwise.
- Specifically, the ability to operate at Y12 and Savannah River as exemplars of different nuclear processing sites is valuable, amplified by the collaboration between the Arkin, Hazen, Alm, and Duff labs.

Conclusion

- It is critical to have a standard protocol for collection and processing of the data.
  ↪ Although it is possible to use data from the global community to help train the model, individual datasets differ in sequencing/extraction/etc. protocols. This needs to be addressed in addition to batch effect.[6]
  ↪ The **metadata** of a site and its history is essential to good data science, and this remains challenging as few standards applicable to this type of study exist.
- However, even with relatively low-resolution measurements, we see strong association of samples from similar contaminated sites clustering across studies.
- To better power our models, further (and higher resolution) measurements from contaminated sites will be added to the dataset.

Acknowledgements

- I would like to thank the members of the Arkin and Hazen Labs for their guidance and support, and the Hazen Lab for their contribution to data collection.

References

This work was funded in-part by the Consortium for Monitoring, Technology, and Verification under DOE-NNSA award number DE-NA0003920