# Assessing the Impact of Nuclear Contamination on Soil Microbial Communities

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

## Introduction

- Nuclear processing (NP) activities often lead to the dissemination of materials and wastes into the surrounding environment.[1]
- The environment undergoes changes as a result of the reciprocal impact between these materials, leading to shifts in the spatial and temporal distribution of microbes around the material source.
  - The contaminants involved (e.g., heavy metals, nitrate, and other organics) can change which microbes can thrive and survive.[2]

**Microbial community variation around sites of known nuclear contamination history can train a machine learning model to predict the contamination source's type, age, and distance.**

## Materials and Methods

- The Earth Microbiome Project (EMP) contains over 2,000 16S sequencing data from soil samples across the globe.[3]
- Using the EMP dataset as a background can help minimize the false positive rate (FPR).
- Additionally, the dataset includes nuclear-contaminated soil data obtained from MTV collaborators and publicly available data from the European Nucleotide Archive (ENA)[4] (Fig. 1).

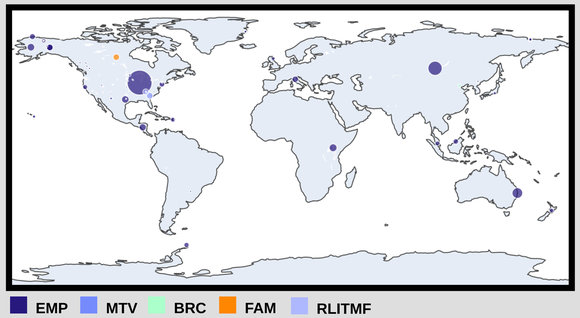


***Fig. 1: Map of soil sample locations used in the analyses.***

- It is crucial to process all the data using the same common workflow (Fig. 2).[5]

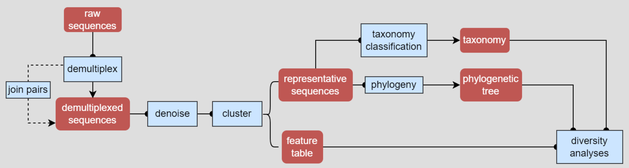


***Fig. 2: QIIME2[6] workflow.***

## 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 + Discussion

- Analysis of composition of microbiomes (ANCOM) analysis (Fig. 3) can be used to compare the composition of microbiomes in 2+ populations and identify significantly different taxa.[6]
  - 484 taxa (at the genus level) were considered to be significantly different between the nuclear-contaminated and background samples.

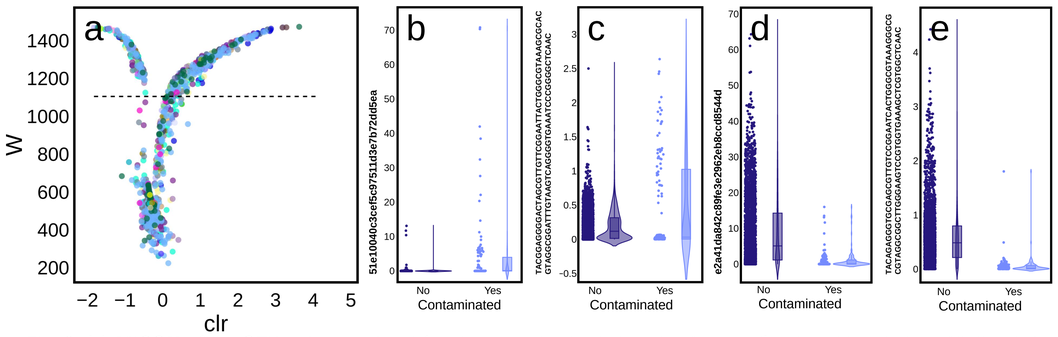


***Fig. 3: (a) ANCOM of the OTUs at genus level (nuclear- contaminated vs. background) colored by phylum. Violin plots of relative abundances of selected taxa that were (b)-(c) more or (d)-(e) less abundant in the contaminated samples.***

- Principal Component Analysis (PCA) is a dimensionality-reduction method (Fig. 4a) that can increase the interpretability of large datasets while minimizing information loss.[7]
  - Each PC captures very little variation from the data.
- Principal Coordinate Analysis (PCoA) is a statistical method (Fig. 4b) that converts a matrix of distances between samples to a map of the dimensions that account for the maximum distances.[8]
  - Two clusters of contaminated samples differ greatly; separations of contaminated samples from background samples are apparent.

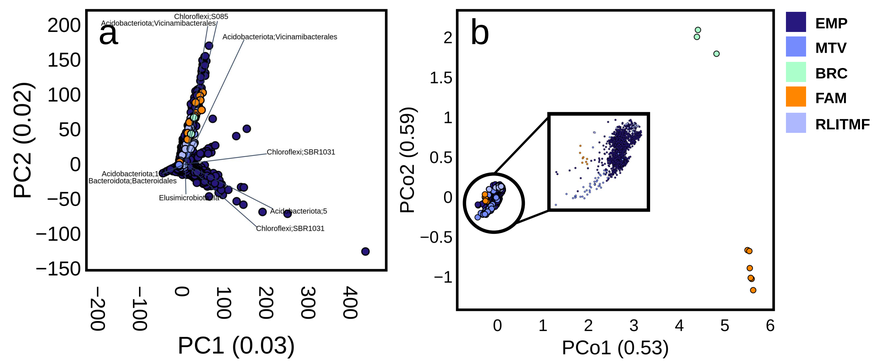


***Fig. 4: (a) PCA score plot overlaid with loadings of the most important features in PC1 and PC2; (b) PCoA plot (with closer view of the left cluster embedded); each point represents a sample.***

## 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.

## Conclusions + Next Steps

- 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, differences in data type (amplicon vs. metagenomics) and sequencing/extraction protocols can present challenges that we need to address.
  - 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 Lab for their guidance and support, especially David Bernstein, as well as our collaborators from the Hazen Lab for their contribution to data collection.
- This work was funded in-part by the Consortium for Monitoring, Technology, and Verification under DOE-NNSA award number DE-NA0003920.

**References + Interactive Plots + Video Presentation** →