Using metagenome assembled genomes (MAGs) to investigate thermophiles in the soils overlying the Centralia, PA coal fire

Jane Lee[1,2], Jackson Sorensen[3], Ashley Shade[3,4]

[1]The George Washington University, [2]ACRES, [3]Department of Microbiology and Molecular Genetics, Michigan State University, [4]Program in Ecology, Evolutionary Biology and Behavior, Michigan State University

Background

- Centralia, PA is the site of a long burning underground coal fire that increases the temperature of the overlying soils.
- Thermophiles are organisms that require high temperatures to grow but are ubiquitous, though understudied, in temperate soils.
- This study investigates the thermophiles in the soils overlying the Centralia coal fire through metagenome assembled genomes or "MAGs".

Research Question

What is the variability of genome content among thermophile populations, given temperature gradients in Centralia soil?

Objectives

To discover gene loss and gain when comparing thermophile genomes within populations inhabiting Centralia soils overlying the coal mine fire.

Methods

Sampling, DNA Extraction, and Sequencing

Field Sampling
13 sites were sampled, 6 fire affected, 6 recovered, and 1 reference site

Extracting DNA
2 DNA samples were extracted from Cen13: an **uncultured** sample from a soil slurry and a **culture** grown at 60ºC

Sequencing
DNA from all Centralia sites were sequenced using shotgun sequencing with Illumina

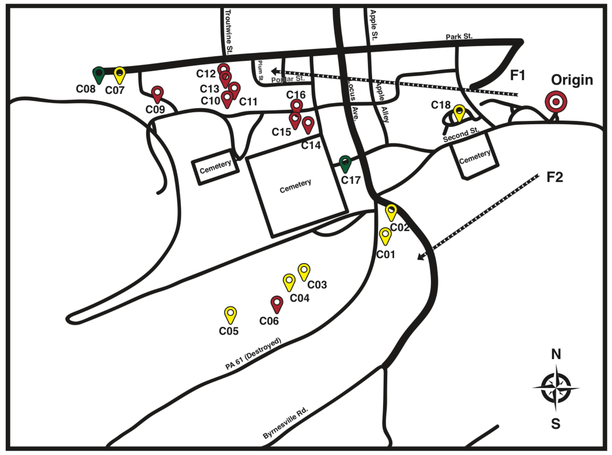

Figure 1. Map of Centralia, PA. Red sites are fire affected, yellow sites recovered, and the green site is the reference site.[1]

Constructing MAGs:

Quality Control
khmer[2] and fastx toolkits[3]

Assembly
Megahit[4]

Mapping
bbmap[5]: aligns reads from other sites to the assembly

Binning
MetaBAT[6]

MAG Analysis:

MAG

Gene Annotation: Prokka[9]

Remapping

Taxonomy: MiGA[7] & CheckM[8]

Bin Statistics: CheckM[8]

Gene loss/gain

Results

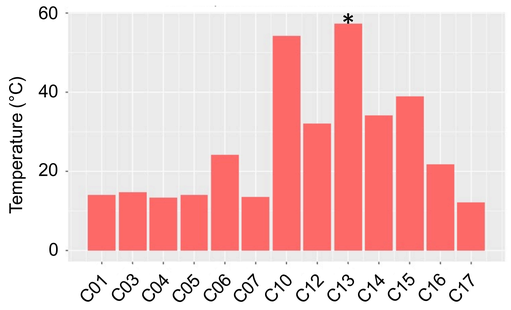

Figure 2. Temperatures (°C) are listed for each site in Centralia.[1] * marks the site that was used for cultivation and assembly.

Assembly Statistics	Uncultured	Cultured
# contigs (>= 0 bp)	508410	60865
# contigs (>= 1000 bp)	222766	28137
Total Length (>= 0 bp)	898424779	117928436
Total Length (>= 1000 bp)	698652595	95212184
# contigs	508410	60865
Largest Contig	413537	322840
Total Length	898424779	117928436
GC (%)	63.9	48.68

Table 1. Assembly statistics from the uncultured and cultured Cen13 metagenomes. The cultured has a much lower GC content than the uncultured assembly from the same sample, suggesting different membership.

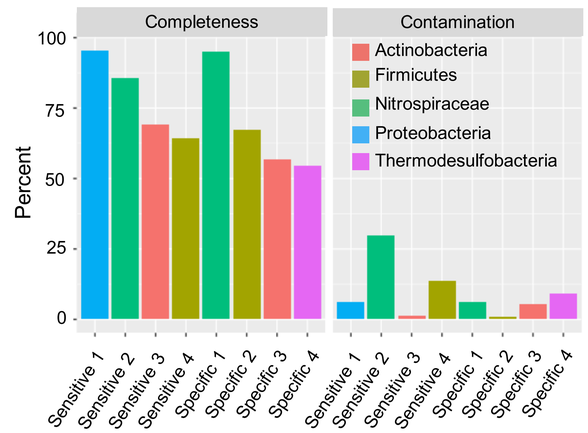

Figure 3. The assembled contigs >2.5 kbp from the uncultured Cen13 DNA were binned using the "very specific" and "very sensitive" settings on MetaBAT. MAGs with >50% completeness are shown. We identified taxonomy of the MAGs by querying them against the NCBI Prokaryote project in MiGA[7].

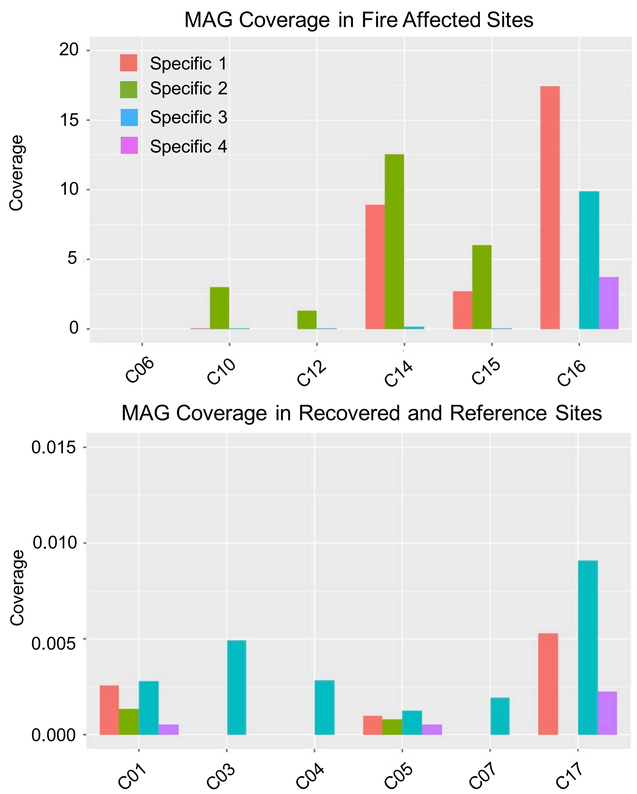

Figure 4. Abundance patterns of MAGs across sites in Centralia, PA. Coverage is defined as the average number of times each base of the MAG was sequenced in the metagenome. Note differences in y-axis ranges. P-values from correlation tests between MAG coverage and temperature were not significant.

Results

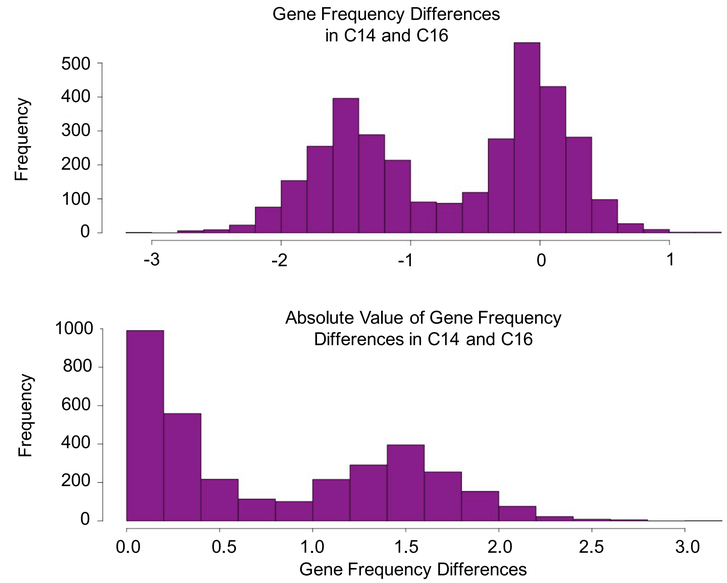

Figure 5. MAG "Specific 1" (Figure 3) was chosen as our target population. C14 and C16 were chosen as sites of interest because they had high MAG gene coverage and were fire affected. MAG gene frequencies within a site were calculated for each site by dividing their coverages by the median gene coverage of all genes in the MAG. Genes with abs(differences) > 1.0 suggest they are in different frequencies across the two sites.

Conclusions

- Nearly complete and minimally contaminated MAGs can be binned from complex soil metagenomes (Figure 3).
- Within a MAG population, genes have different abundances at different sites (Figure 5), suggesting loss or gain.

Project Limitations

- High levels of diversity in soil make it difficult to observe thermophiles in recovered and reference sites with metagenome sequencing because of low coverage.

Future Work

- Construct MAGs from cultured sample Cen13, and interrogate their distribution and diversity.
- Perform gene frequency analysis on other MAGs, and across other sites.
- This will be the 4th years of sampling the soils above the Centralia mine fire, allowing for an analysis of MAG gene content within a site through time.

Acknowledgements

This work was supported by the Joint Genome Institute (CSP-1824), the Institute for Cyber-Enabled Research at Michigan State University, and the NSF Advanced Computational Research Experience for Students (ACRES) REU program funded by grant #1560168. We thank Tammy C. Tobin for helpful discussions.

References

1. Lee, S., Sorensen, J.W., et al. (2017). Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. ISMEJ, 1447–1459.
2. Crusoe, M.R., et al. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research, 4:900.
3. FASTX Toolkit http://hannonlab.cshl.edu/fastx_toolkit/index.html by Hannon Lab
4. Li, D., Liu, C., Luo, R., et al. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31 (10): 1674-1676.
5. http://jgi.doe.gov/data-and-tools/bbtools/
6. Kang, D.D., et al. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 3, e1165.
7. Rodriguez, R. In preparation. Microbial Genomes Atlas: Standardizing genomic and metagenomic analyses for Archaea and Bacteria.
8. Parks, D.H., et al. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25: 1043–1055.
9. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30(14):2068-9.