##### Soil Genome Database Internship

Graduate student mentor: Zoey Werbin ([zrwerbin@bu.edu](mailto:zrwerbin@bu.edu)), Bhatnagar Lab

Running notes document: [link](https://docs.google.com/document/d/1wKl_h48osGW8pGdwH2SGfX-0Vw5trqxMXFdtWXHdVxQ/edit?usp=sharing)

##### Background:

Soil microbes are difficult to culture in lab settings (Bartelme et al. 2020), which has made 'omics approaches a critical route for studying their activity. However, the lack of cultured isolates also means that most soil taxa have not had their genomes sequenced (Nesme et al. 2016). Recently advances in bioinformatics have enabled the assembly of soil genomes directly from short-read metagenome sequences (called Metagenome-Assembled-Genomes, or MAGs). Genome databases often lack up-to-date MAGs due to their rapid publication rate (Youngblut et al 2021), and MAGs are frequently missing the metadata that would allow researchers to have confidence in their quality, such as contamination and completeness scores (Breitweiser et al. 2019; Bowers et al. 2017). Even with inclusion of MAGs, the NCBI Genome database is overwhelmingly biased toward medically-relevant microbes. Taken together, these data issues limit our ability to characterize the DNA found in soil samples (Werbin et al. 2022). Soil ecology research would therefore benefit from a specialized up-to-date repository of microbial taxa.

**Project goal:**

Develop a flexible database of genomes associated with soil fungi, bacteria, archaea, and viruses, with associated code for updating with genomes upon publication.

To classify the taxonomy of short reads from metagenomic projects, Kraken2 databases are often built using subsets of the ~200k genomes indexed by RefSeq. Database creation with novel genomes requires custom formatting of taxonomy and sequence data, and filtering of low-quality or contaminated genomes. This process has been implemented in a new pipeline called [Struo2](https://peerj.com/articles/12198.pdf) (Youngblut and Ley, 2021), which allows new genomes to be added to databases created from the Genome Taxonomy DataBase (GTDB). If possible, we want to filter the GTDB to remove the non-soil taxa.

To identify improvement in the percentage of identified reads, results can be compared with a pre-built RefSeq-based database such as PlusPF.

**Key readings:**

[Parks et al 2022](https://academic.oup.com/nar/article/50/D1/D785/6370255) - GTDB paper

Bowers et al. 2017 - describes the benchmarks for what makes a MAG "good" or "bad"

Youngblut and Ley, 2021 - [Struo2](https://peerj.com/articles/12198.pdf) pipeline that we'll use as a baseline to work from

**Additional resources:**

[Script](https://github.com/hcdenbakker/GTDB_Kraken/blob/master/scripts/gtdbToTaxonomy.pl) for converting GTDB taxonomy to Kraken2 format

[Another example](https://github.com/bhattlab/kraken2_classification/blob/master/manual/mag_db.md) of how the novel genome database can be created, with a rough sketch of including taxonomy.

[Checking](https://github.com/DerrickWood/kraken2/issues/143) if a new genome was successfully added to Kraken2 database

[Explanation](https://github.com/DerrickWood/kraken2/issues/240) of nodes.dmp:

Visualization: [Pavian](https://fbreitwieser.shinyapps.io/pavian/) shiny app

[Struo2 script](https://github.com/leylabmpi/Struo2/blob/master/bin/db_create/kraken2/Snakefile) for creating Kraken2 database

## **Soil genomes/genome collections:**

**JGI GOLD (genomes online database) -** allows for searching by soil-associated organisms.

Previous GOLD Search (Oct 2021) downloaded to the SCC at:

/projectnb/microbiome/zrwerbin/NCBI\_bacterial\_genome\_list.csv (7500 genomes)

/projectnb/microbiome/zrwerbin/NCBI\_fungal\_genome\_list.csv (Mycocosm)

**However, the NCBI taxon ID doesn't seem to work quite right.**

**Bacterial/archaeal genomes**

RefSoil: 922 genomes from NCBI, as of Feb 2016 (ID in Table S1, [Choi et al. 2017](https://doi.org/10.1038/ismej.2016.168))

294 MAGs: Nayfach et al. 2020

MAGs from NEON

Unclear if [TerraGenome/MetaSoil project](https://www.fao.org/agriculture/crops/thematic-sitemap/theme/spi/soil-biodiversity/research-into-soil-biodiversity/the-terragenome-project/en/) ever published genomes?

67 soil MAGs: ​​Alteio et al. 2020

**Fungal genomes**

Mycocosm

**Viral genomes**

2961 viral OTUs: [Santos-Medellin et al. 2021](https://github.com/cmsantosm/SpatioTemporalViromes/)

1910 viral OTUs: [Lee et al. 2021](https://www.biorxiv.org/content/10.1101/2021.10.20.465127v1.full.pdf) JGI Genome Portal (DOI: 10.25585/1487501)

Possibly more from Hazard Lab ?

Soil Plasmids: [RefSoil+](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6392096/#B21)

**Associated project:**

We want to maximize the database coverage for a specific environment (temperate forest soils) by targeted assembly of ribosomal RNA genes, which increase the number of genomes that pass the quality cutoff for inclusion.

Input: raw metagenome short reads from Harvard Forest and Bartlett soils. Pipeline first sorts reads that are suspected as rRNA genes, then assembles those reads, then we match to available genomes.

<https://github.com/leylabmpi/ResMiCo/tree/master/ResMiCo-SM>

<https://github.com/leylabmpi/ResMiCo/blob/master/notebooks/01_resmico-sm/02_GTDBr89_n9k-rand/01_train.ipynb>

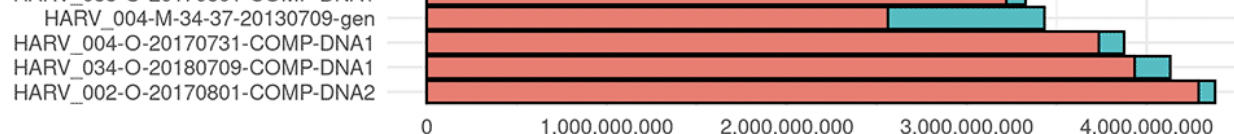
Test samples:

3-4 samples from Harvard Forest

/projectnb/talbot-lab-data/metabolic\_models/scripts/metaGEM/dataset/

/projectnb2/talbot-lab-data/metabolic\_models/scripts/metaGEM/qfiltered/HARV\_034-O-20190718-COMP-DNA1/HARV\_034-O-20190718-COMP-DNA1\_R2.fastq.gz

/projectnb2/talbot-lab-data/metabolic\_models/scripts/metaGEM/qfiltered/HARV\_034-O-20190718-COMP-DNA1/HARV\_034-O-20190718-COMP-DNA1\_R1.fastq.gz

****

**Works Cited**

Bartelme, R. P., Custer, J. M., Dupont, C. L., Espinoza, J. L., Torralba, M., Khalili, B., & Carini, P. (2020). Influence of substrate concentration on the culturability of heterotrophic soil microbes isolated by high-throughput dilution-to-extinction cultivation. Msphere, 5(1), e00024-20.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35(8):725–31.

Florian P Breitwieser, Jennifer Lu, Steven L Salzberg, A review of methods and databases for metagenomic classification and assembly, Briefings in Bioinformatics, Volume 20, Issue 4, July 2019, Pages 1125–1136, <https://doi.org/10.1093/bib/bbx120>

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth’s microbiomes. Nat Biotechnol [Internet]. 2020 Nov 9; Available from: http://www.nature.com/articles/s41587-020-0718-6

Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, et al. Back to the future of soil metagenomics. Front Microbiol. 2016;7(FEB):1–5.

Santos-Medellin, C., Zinke, L.A., ter Horst, A.M. et al. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. ISME J 15, 1956–1970 (2021). https://doi.org/10.1038/s41396-021-00897-y

Werbin ZR, Hackos B, Lopez-Nava J et al. The National Ecological Observatory Network’s soil metagenomes: assembly and basic analysis [version 2; peer review: 2 approved with reservations]. F1000Research 2022, 10:299 (https://doi.org/10.12688/f1000research.51494.2)