PROBLEM STATEMENT – Soil microbes are difficult to culture in lab settings (Bartelme et al. 2020), and the lack of cultured isolates means that most soil taxa have not had their genomes sequenced (Nesme et al. 2016). Recent 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 genomic information for microbial taxa.

PROJECT GOAL – Student will collaborate with PhD candidate Zoey Werbin to generate a Kraken2 database of soil microbe genomes using the Struo2 pipeline, which is still in active development. This will involve (1) familiarization with the Struo2 pipeline and with the use of Kraken2 for read classification, (2) wrangling of genome metadata from a variety of sources, including the GTDB, the JGI’s GOLD, and assorted publications, and (3) application of the Struo2 pipeline for database generation and updating. The effectiveness of the resultant Kraken2 database for soil microbe genome sequencing read classification will be compared with benchmarks associated with existing databases. Once the performance of the database has been optimized, preparations will be made to publish results and construct a web interface with which researchers may use the database for read classification.

STUDENT CONTRIBUTION THUS FAR – Student has downloaded and set up the Struo2 pipeline, successfully running its Kraken2 database creation and database updating functions on creator-provided test datasets. Student has analyzed, filtered, and processed sample metadata for genomes in the Genome Taxonomy Database (GTDB) and the Genomes Online Database (GOLD) associated with the Joint Genome Institute (JGI). Student has created and maintained a GitHub repository to contain all project-associated R scripts for data analysis and processing.

STUDENT OBJECTIVES MOVING FORWARD – Student will continue to troubleshoot the use of the Struo2 pipeline to update a downloaded, pre-made Kraken2 database of genomes found in the GTDB with selected genomes from JGI GOLD. Student will analyze, filter, and process additional genome metadata from a variety of literature sources and add those genomes to the growing database. Student will collaborate with Zoey Werbin on the writing of a paper reporting the database’s construction and effectiveness in classifying soil microbe genome sequencing reads.

PROFESSIONAL DEVELOPMENT IMPACT

This project will provide the student with added experience in the use of two critical components of a bioinformatician’s toolkit: command line programming and workflow management systems like Snakemake. This is of particular value to the student in question, as his skill level with Linux lags behind his skill level in R, Python, SQL, and web development, and he has had minimal prior experience with Snakemake or other workflow management systems.