Gene-finder pipeline

1. Curate a reference sequence database for a gene of interest and build a multiple sequence alignment.
2. Run Snakemake **gene-finder** to perform **HMMER** searches against metagenomes with the reference sequence alignment
3. Jupyter notebook to analyze outputs from gene-finder and select putative gene sequence hits (e.g. NifH.ipynb)
4. Rebuild reference sequence alignment including sequences collected from 3 and rerun Snakemake **gene-finder** a second time
5. Jupyter notebook to analyze second round outputs from gene-finder and collect additional gene sequence hits (e.g. nifH_rnd2.ipynb)
6. Jupyter notebook to combine putative **unique gene sequences from both rounds** (e.g. nifH_combo.ipynb)

Gene abundances

1. Use **Diamond** to map unique gene sequences (clustered at 98% identity) against all FastQ reads
2. Jupyter notebook to parse diamond output files to compute gene sequence abundance within the metagenomes or metatranscriptomes, normalized for multimapping reads, sequencing depth of the associated sample run, and length of the specific protein sequence for gene abundance in RPKM (nifH_diamond_parse.ipynb)

Plot gene abundances globally and by depth profile as well as environmental variables using Cartopy (e.g. cartopy_abundance_analysis_simple.ipynb)

MAG-analysis

1. Jupyter notebook to associate the gene sequences with MAGs in which they are found (e.g. nifH_MAGs.ipynb)
2. Jupyter notebook to create a summary of all MAGs which contain one of the genes of interest (e.g. MAGs_analysis.ipynb). Only if performing analysis on multiple genes in parallel.
3. Perform a **fastANI** analysis on the MAGs of interest.
4. Run **CheckM** and **GTDB-TK** or other quality and taxonomy predictions
5. Jupyter notebook to analyze fastANI output; cluster based on 95% similarity and overlay CheckM and GTDB-TK data (e.g. nifH_fastani_2_9.ipynb).
6. Run Snakemake kobra-annotation for automated **KEGG annotation** on MAGs of interest.

Visualize clusters using **Anvi'o pangenomics**

Visualize presence and absence of genes of interest among MAGs by parsing the KEGG annotation tables (e.g. nifH_blast_koala_2_7.ipynb)