

Scripts

`genbank_grab.py` is a Python 2.7 script that reads in a comma-separated file containing the genbank accession and other details about phage genomes. It downloads the corresponding genome files in GenBank (full) format and stores them in the corresponding 'genbank' directory under 'genomes', and as FASTA files in the corresponding 'fasta' directory. The script goes through the GenBank annotated features and systematically assigns identical `phage_##id`'s to both `/gene` and `/CDS` features, for both `/gene` and `/locus_tag` qualifiers. This is done so that PhamDB will process the records in a coherent way, giving this same `phage_##` identifier to all genes in all genomes, so that when a pham table is exported through Phamerator it will consistently align with the names from the genbank record, so that they can be mapped back. The script writes out the input table as 'Processed_table.csv', adding metadata available in GenBank annotations (e.g. genome length, host...).

`Genome_parsing.py` is a library used by `genbank_grab.py` that contains a set of functions to extract specific records from genome files and save them in specific formats.

`get_genomes.py` is a Python script to obtain phage and host genomes in GenBank format. Phage genome GenBank files are obtained from a locally downloaded phage genome database. Host genome GenBank files are copied from previous analyses done.

`file_for_plot.py` is a Python script used to construct the final csv files required to obtain the plots with `ggplot2`.

`plot_with_R.txt` is an R script used to construct the plots with `ggplot2`.