CONSIDER FOR TALK

2021 SEA Faculty Meeting Abstract

University of Nevada Las Vegas

Las Vegas NV

Corresponding Faculty Member: Philippos Tsourkas (philippos.tsourkas@unlv.edu)

Phage Commander, an app for rapid phage genome annotation using multiple programs

Matt Lazeroff, Geordie Ryder, Sarah Harris, Philippos K Tsourkas

Genes in newly sequenced phage genomes are usually identified using one or more freely-available gene identification programs (e.g., Glimmer, GeneMark family, RAST, Prodigal, etc.). Each program uses a distinct algorithm and produces unique results. No program has been shown to consistently outperform the others; thus, the choice of which program to use is not obvious. Using multiple programs improves accuracy, as multiple programs may identify genes missed by an individual program, and we can decide which genes to keep by consensus. However, this requires integrating the results of multiple programs and increases the labor cost of annotation. DNA Master, for example, conveniently bundles Glimmer and GeneMark within its user interface; however if we wish to use an additional program (e.g. host-trained GeneMark.hmm), the program has to be run outside DNA Master and the genes it identifies spliced in manually.

We present an app developed in our lab named Phage Commander, which runs a phage genome sequence through ten gene identification programs simultaneously (Glimmer, GeneMark, host-trained GeneMark.hmm, GeneMark S, GeneMark S2, GeneMark with Heuristics, RAST, Prodigal, MetaGene, and Aragorn), and outputs the results in an easy-to-visualize spreadsheet, where each row is a gene. Rows are shaded based on how many programs identify a particular gene, with darker shading corresponding to more programs identifying a gene (i.e., a stronger call).

Phage Commander can export its output as an Excel spreadsheet, or as properly formatted output files (.gb), which can be easily imported into DNA Master for further processing. Users can select the threshold for determining which genes should be exported (e.g. genes identified by at least one program, genes identified by at least two programs, etc.). Phage Commander was benchmarked using eight high-quality phage genomes whose genes are backed by experimental data. Results show that the most accurate annotations are usually obtained by exporting genes identified by at least two programs (not counting Aragorn); this setting typically minimizes the sum of false positives (non-genes) and false negatives (missed genes).

Phage Commander is currently being used in our Phage Discovery class at UNLV and has significantly reduced the labor cost of genome annotation and likelihood of errors. Phage Commander is freely available for download from GitHub (https://github.com/sarah-harris/PhageCommander) as an easy-to-use, one-click executable file and runs, on Mac, Windows, and Linux. A manuscript describing Phage Commander is currently in press at the journal PHAGE: Therapy, Applications, and Research.