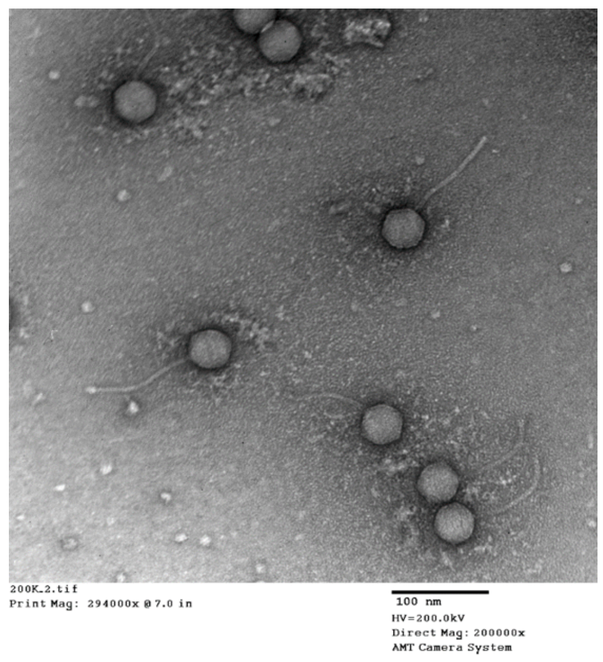
GENE ANNOTATION OF MYCOBACTERIOPHAGE KADY USING *IN SILICO* GENE IDENTIFICATION METHODS

Emily K. Gorkiewicz , Christina E. Jones, Jacob D. Kagey and Joshua J. Thomson

Abstract

Effective treatment of *Mycobacterium tuberculosis*, a deadly pathogen affecting millions of people worldwide, has proven to be difficult due to the increasing prevalence of multiple antibiotic-resistant strains. In an effort to solve this global health concern, researchers are beginning to turn to an alternative approach known as phage therapy, which uses host-specific bacterial viruses (bacteriophages) to kill the intended pathogenic bacteria without compromising the health of the infected human. Kady is a novel mycobacteriophage collected from a soil sample in southeastern Michigan and isolated using *Mycobacterium smegmatis* as the host. The isolated phage was submitted for genome sequencing and found to be composed of 50,898 bps. Kady was then auto-annotated using the program DNA Master. In this study, we evaluate the accuracy of DNA Master by comparing the predicted genes found to data collected from various bioinformatics tools, namely GeneMark, Glimmer, and Starterator. Each predicted gene was also compared to previously sequenced bacteriophages to determine similarity using BLASTp data and PhagesDB.org. The primary goal of this analysis is to provide further knowledge of the ubiquitous nature of mycobacteriophages in an effort to develop therapeutic uses for treating human diseases such as *M. tuberculosis*.



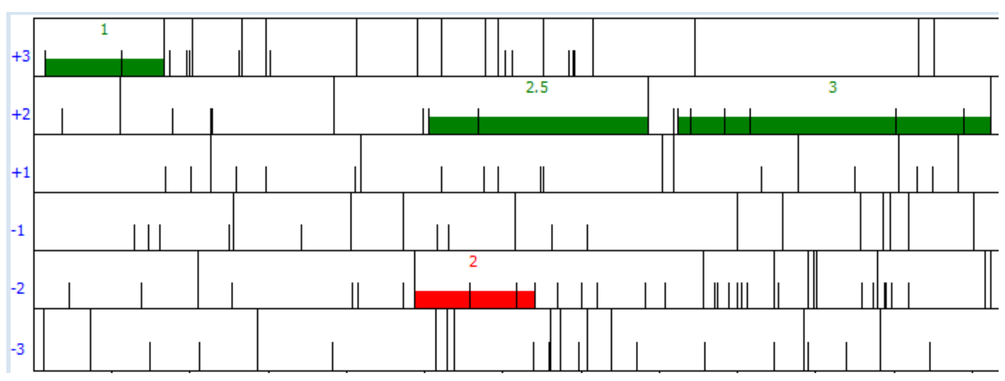


Fig.1 Gp 2 (REV) was deleted in DNA Master and replaced with Gp 2.5 (FWD)

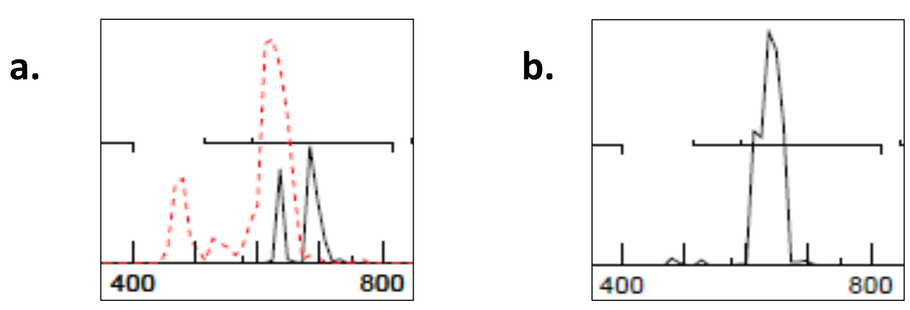


Fig.2 (a) Genemark (b) Glimmer both showing better coding potential for Gp 2.5. Gp 2.5 was added and Gp 2 was deleted due to the lacking potential that was shown.

Introduction

After Kady was auto-annotated by DNA Master, researchers sorted through the predicted genes to determine the accuracy of the program. This was done using:

- Glimmer and Genemark, two programs that use algorithms to predict coding potential of genes in various types of genomes.
- PhagesDB.org BLAST results, which compared the genome of Kady to previously sequenced bacteriophages.
- Starterator, a tool designed to predict the start codon of a given gene when other data conflicts.

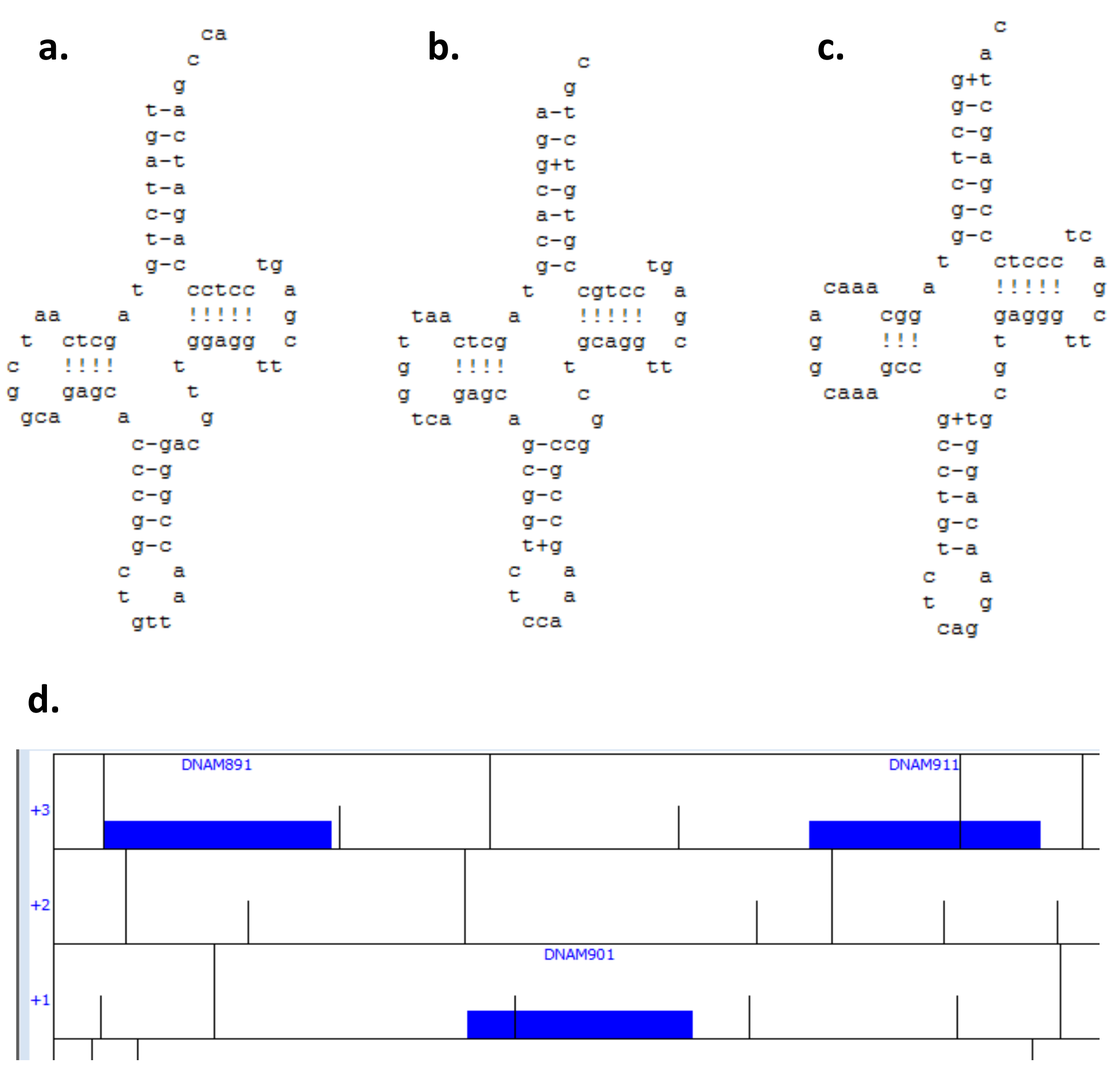


Fig 3 (a) tRNA-Asparagine (b) tRNA-Tryptophan (c) tRNA-Leucine (d) DNA Master Frame view of of the three tRNA found within Kady. (tRNA figures a-c were taken from web-based program Aragorn)

Results and Discussion

We analyzed the accuracy of DNA Master for predicted genes 1-16 and 34-60 of Kady. While a majority of the predictions were correct, the following significant changes were made:

- Gp 2 was deleted because it did not follow the typical arrangement of DNA sequencing and lacked coding potential.
- Gp 2.5 was created to eliminate the significant gap between Gp 1 and Gp 3. The coding potential, as determined by GeneMark (Fig.2a) and Glimmer (Fig. 2b), confirmed the coding potential of this gene.
- The 5' and 3' ends of the tRNAs predicted were trimmed to align with the web-based program Aragorn (Fg.3)

Acknowledgments

This work was funded by the National Institute of Health ReBUILDetroit grant and the Howard Hughes Medical Institute SEA-PHAGES Program. Dr. Joshua Thomson, Dr. Jacob Kagey, and Dr. Chinyere Knight contributed considerable time and effort in educating and advising us over the duration of this project. We would like to thank all parties involved for this opportunity and learning experience.