Metagenomic Exploration the Sequel: Development of novel tools for viral and bacterial sequence analysis

Cody Glickman

CPBS Update Talk

Nov 12th, 2018

Research Update

Clinical NTM Gene Databases

Submitted ... https://mra.asm.org/latest

Duobiome: 18S/16S Parallel Analysis

In progress

Hybrid Viral Contig Prediction

In progress

Virulence Factors in Bacteriophages

Submitted ...

Progress of Other Projects

Asthma Environmental Microbiome

Submitted abstract to ATS

Building Up Domains: Lysogenic Host Discovery

Incorporated into large collaborative NCBI initative

Genomic Retrieval and Blast Database Creation

Accepted Poster ISME 2017

Hawaiian Soil Chemistry and Culture

Submitted ...

Nontuberculous Mycobacterial (NTM) Infections

Number of Cases

The number of NTM cases is estimated over 100K

Increasing Case

The rate of cases is estimated to grow at 8% every year

Populations at risk of developing NTM

- Immunocompromised individuals
- Patients with lung damage or malfunction
- Residents of warm costal areas especially Hawaii

Strollo SE, et al. Ann Am Thorac Soc. 2015
Adjemian J, et al. Am J Respir Crit Care Med. 2012

Laboratory Research Methods

Conditions for NTM Environmental Growth

Identifying important characteristics for NTM growth

Environmental Microbiome

Developing methods to characterize home environments

Clinical NTM

- Developing resources to study clinical NTM
- Identifying potential mechanisms of NTM transmission

Viral Focus

Bacteriophages (Phages)

Phages are DNA viruses that infect prokaryotes

Phage Diversity

Investigating how phage abundance and diversity affect susceptibility to NTM lung infection

Phage Vectors

Researching how phages act as carriers of bacterial genes within clinical NTM infections

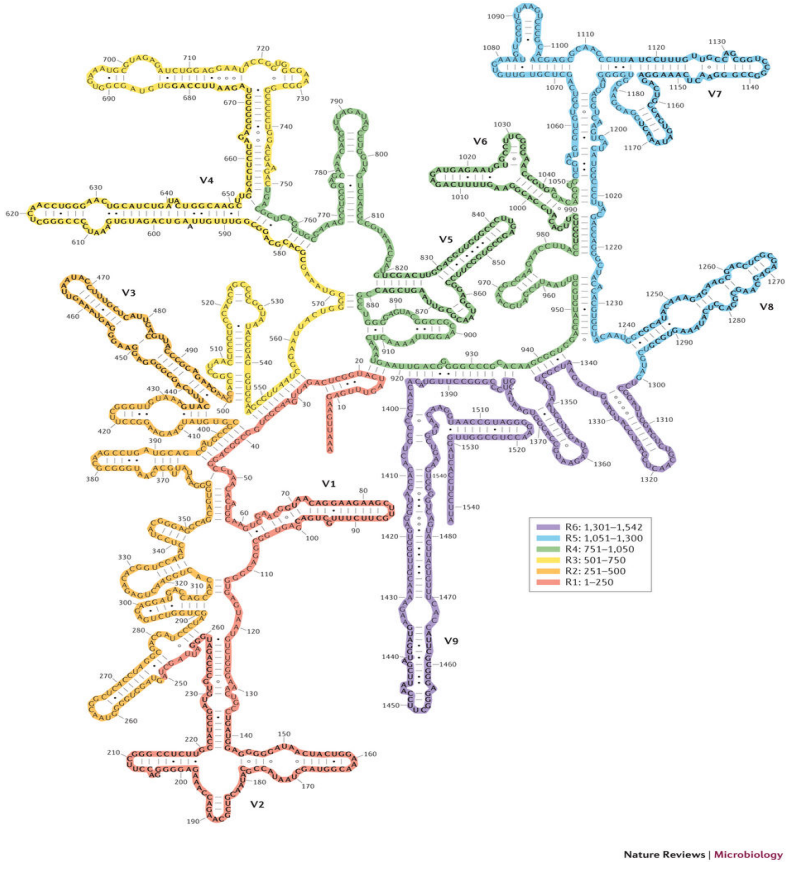
Molecular Methods to Study Phages

Difficulties of phage study

- Lack of universal marker gene
- Sequence heterogeneity
- Misclassification in databases

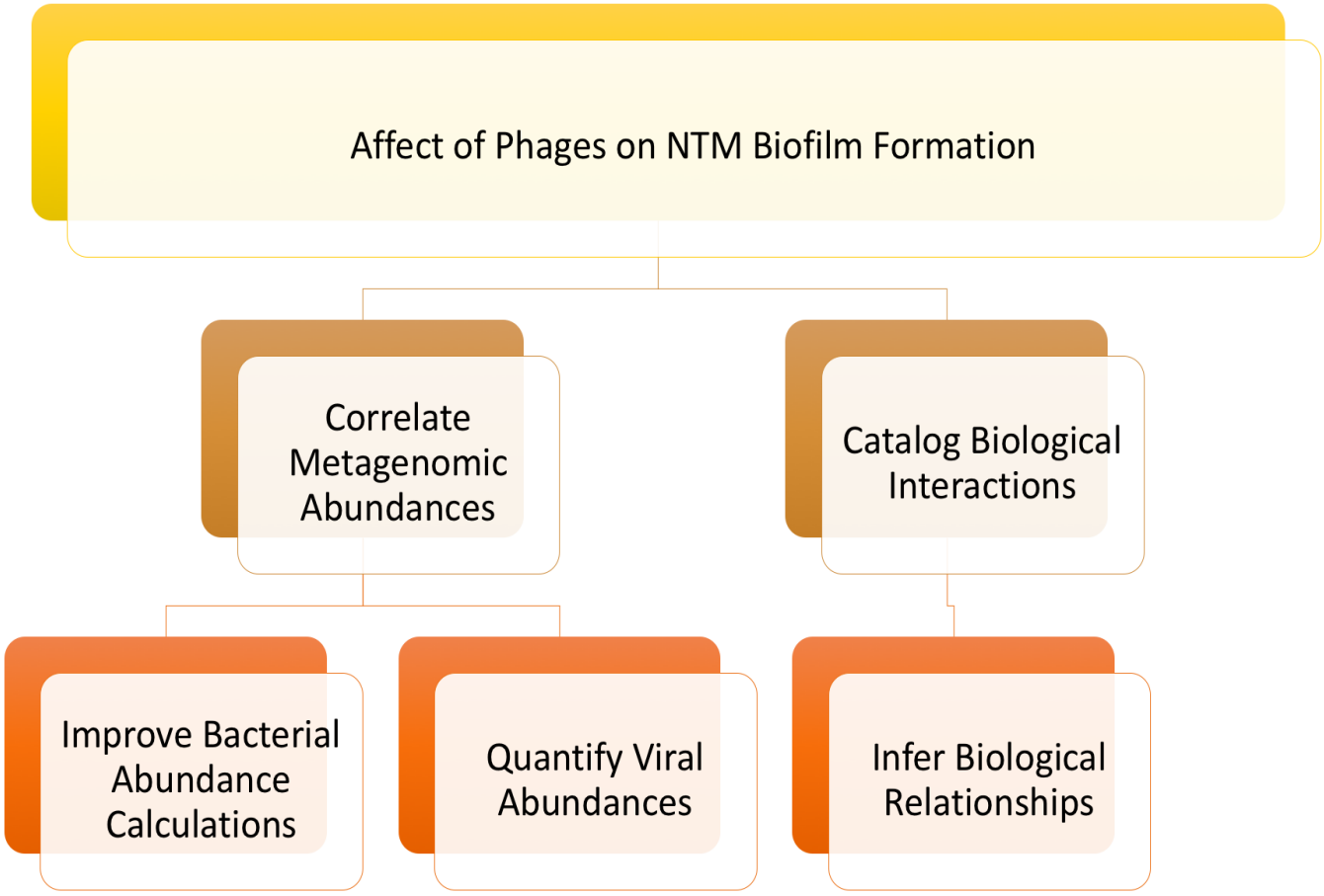
Phage Isolation Methods

- Biological filtration
- In silico methods



Yarza, P., et al. Nature Reviews Microbiology 2014

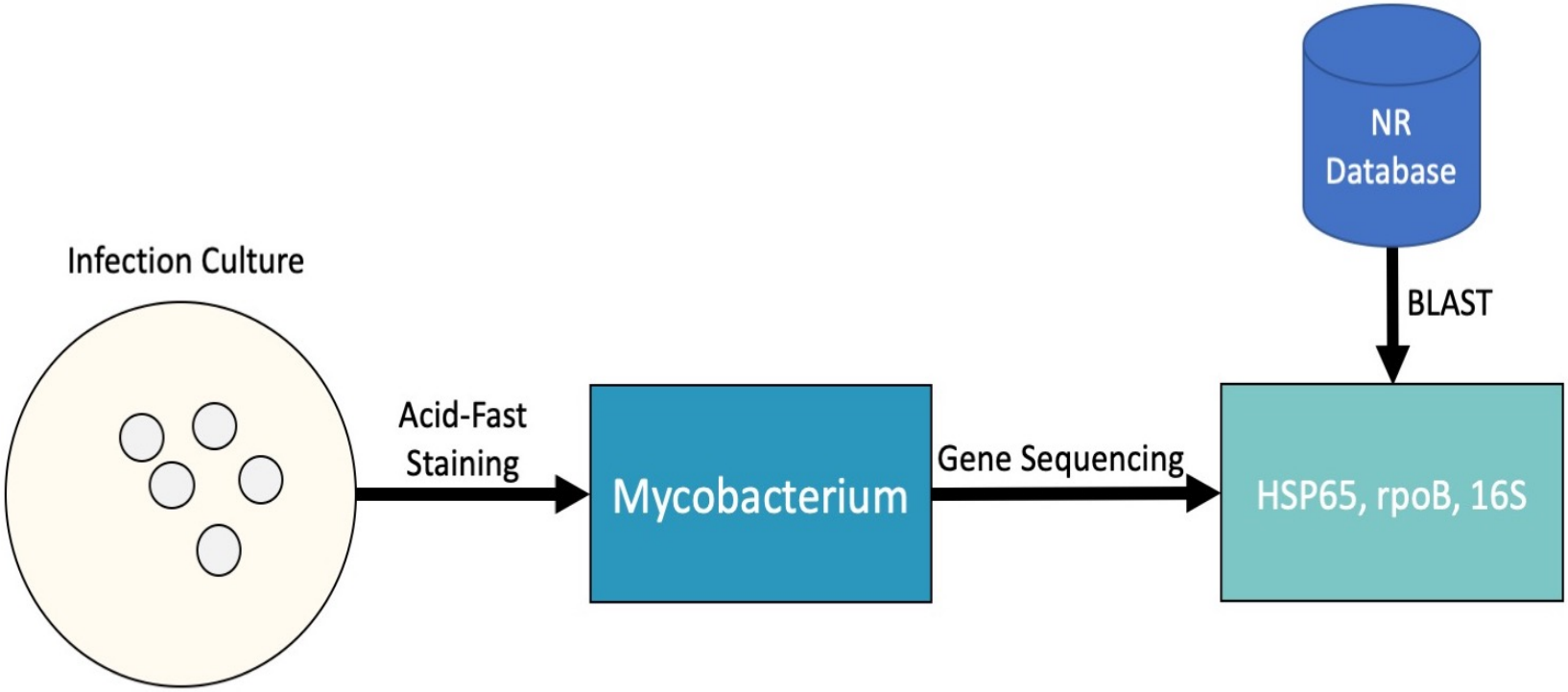
Objective (EDIT)



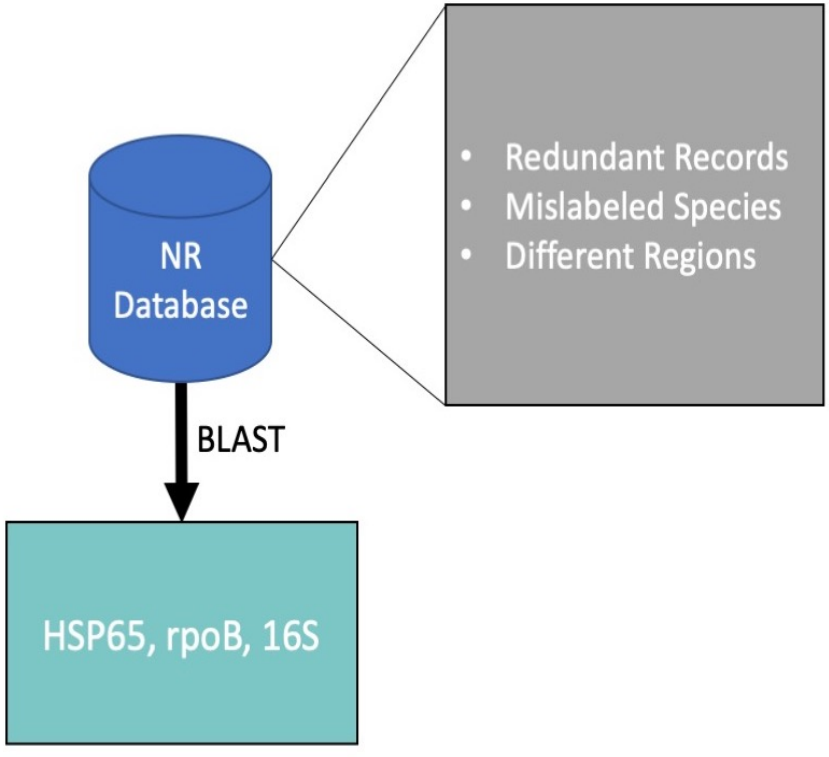
Species Identification of NTM at NJH

Clinical NTM Gene Database

Developed updated database to characterize clinical NTM



Limitations of Current Methods



Redundant Records

Sequences between species are indistinguishable at gene

Mislabeled Species

Naming conventions are constantly updated

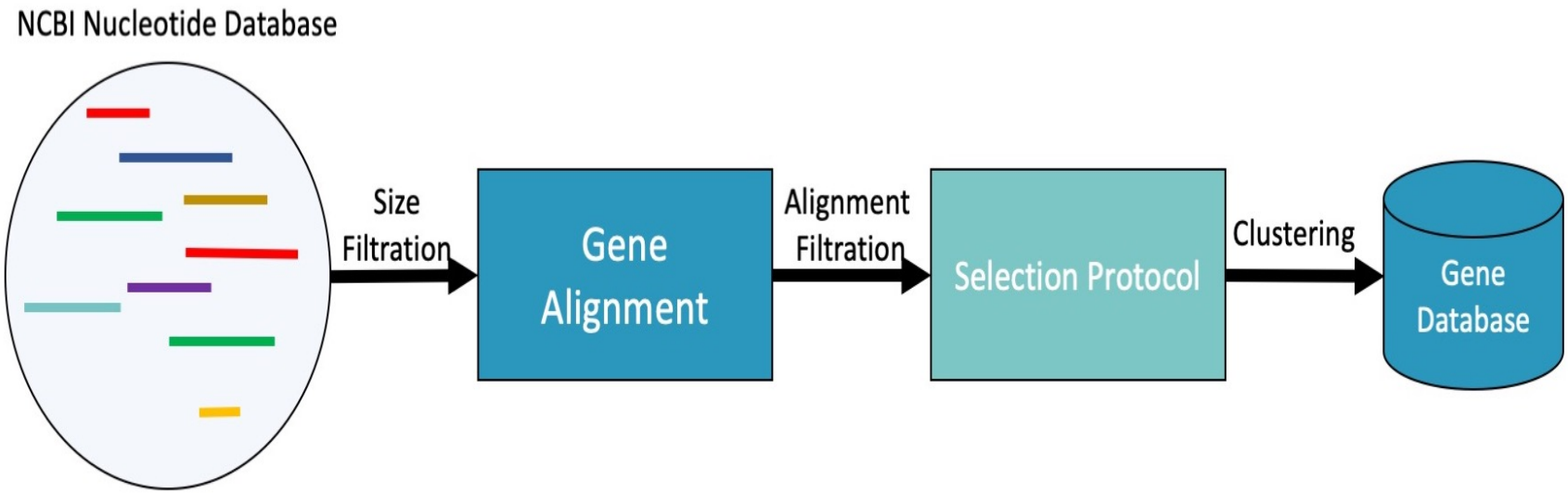
Different Regions

Current protocols amplify specific region of gene

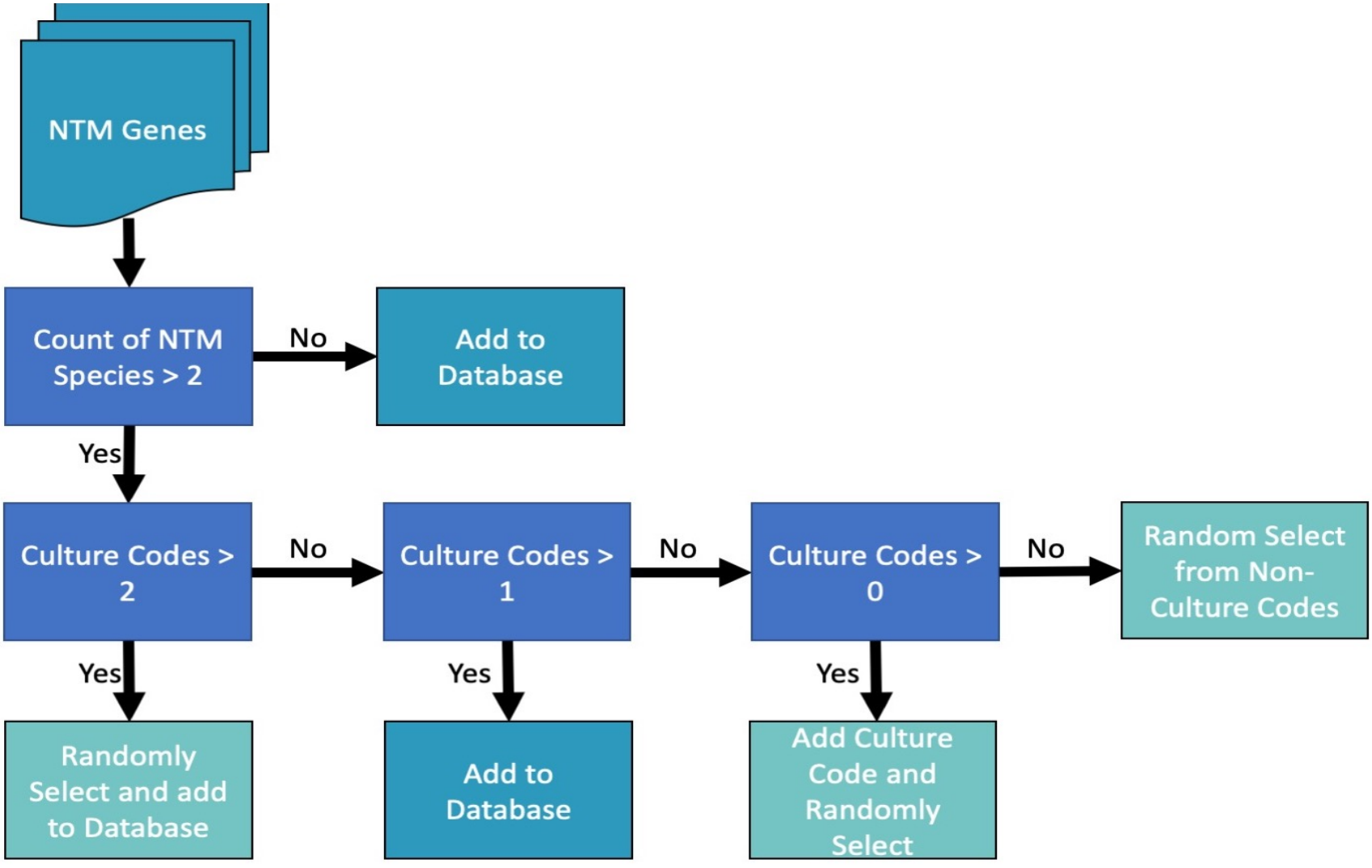
Curated Gene Databases

Number of Sequences per Species

The maximum number of sequences per species in the database is two



Selection Protocol



Clinical Gene Databases

Gene	Region Size	Unique Species
hsp65	382 bases	185
rpoB	657 bases	134
16s rRNA	1470 bases	184

Table: Table 1 highlights the regions lengths and size of the respective databases

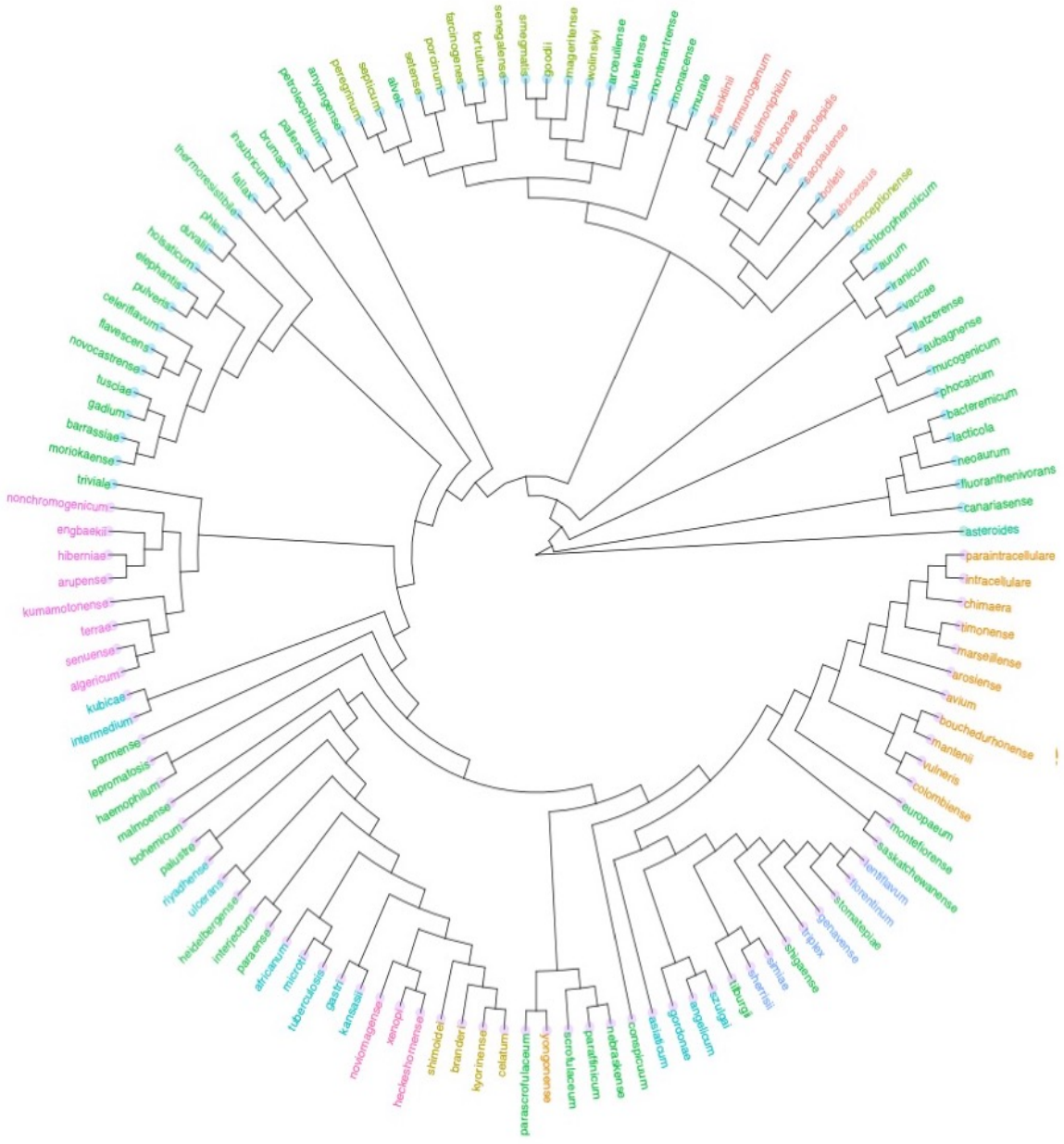
Database Validation

hsp65

154 Species of HSP65 Validation against Subsetted Database
96.73% identical match 5 non matches - two hits in top 5 - two hits not in database (outdated names?)

Dai, J, et al. J Clin Microbiol. 2011

rpoB-hsp65 Tree



Conclusions and Future Directions

Representation

The subsetted database is highly representative of prior published works

Benefits of Curated Database

- Aligned sequences to shared region
- Preferentially selected established culture codes
- Condensed and explicitly labeled ambiguous sequences

Limitations

Size of the gene sequence databases may not differentiate between species

Dai, J, et al. J Clin Microbiol. 2011

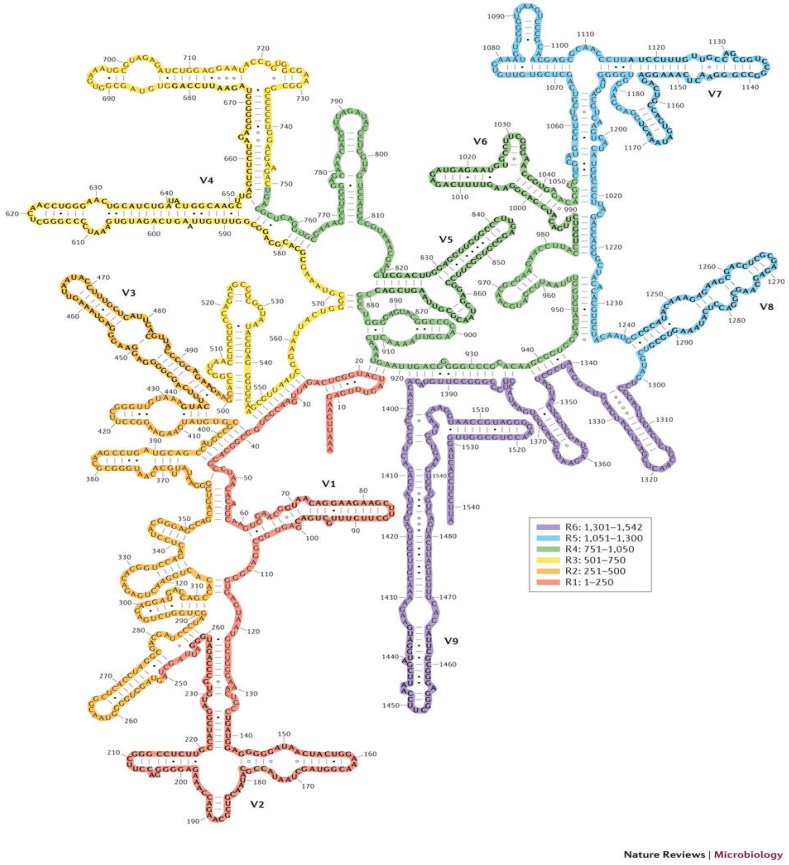
Microbiome

16S Ribosomal RNA Sequencing

- Amplifies a region of gene
- Community level analysis

Traditional Limitations

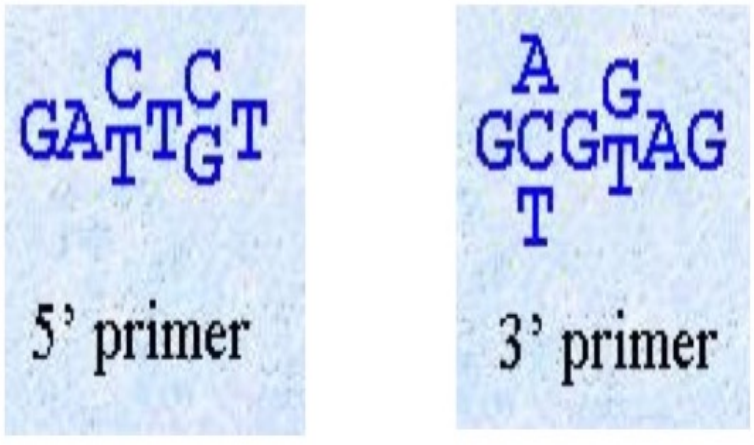
- Biological filtration
- In silico methods



Yarza, P., et al. Nature Reviews Microbiology 2014

Degenerate Primers

Degenerate Primer Example



Caporaso, J.G., et al. PNAS 2011
Wang, Y., et al. PLOS One 2014

Feature of Degenerate Primers

Dual amplification of eukaryotic (18S) and prokaryote (16S)

Universal 16S/18S Primer

515F - 806R primer

Analyze both

Testing against BLAST based and traditional pipeline

Future Directions

Webserver

Shiny web application in development

Viral GRAB

- Expansion of GRAB to viral elements
- Features include ability to filter viruses by genetic material type

Virulence

Virulence Defined

The capacity of a microorganism to proliferate despite host defenses

Influences on Virulence

- Number of microorganisms
- Composition of the mobile genetic reservoir
- Location of niche
- Host immune capabilities

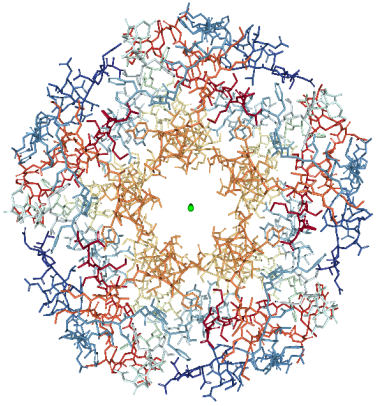
Bacterial Virulence Factors Increase Pathogenesis

Examples of Virulence Factors

- Increased fitness for nutrients
- Host immunity resistance
- Toxin secretion

Diseases from Virulence Factors

Cholera, dysentery, botulism, and food poisoning



PDB Structure of Cholera Toxin

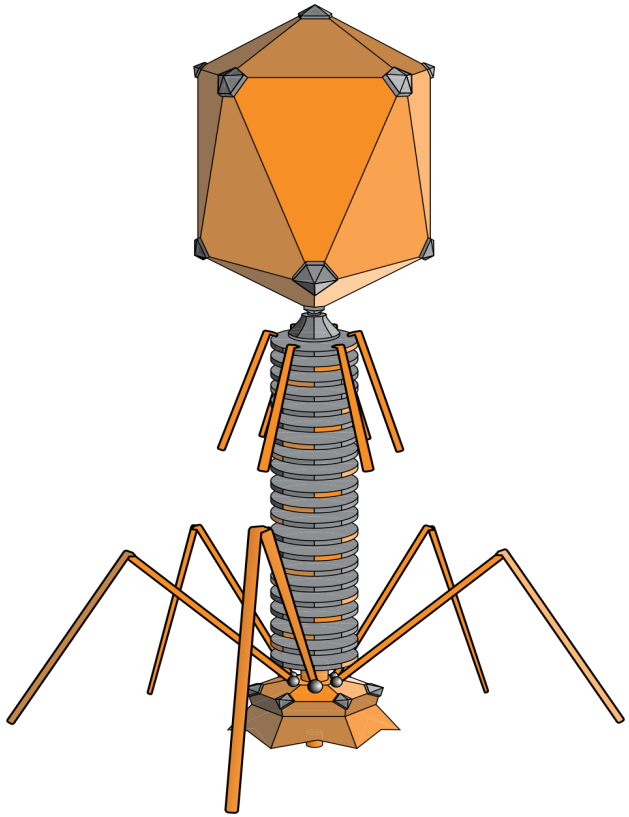
Bacteriophages as a Genetic Reservoir of Virulence Factors Genes

Bacteriophages (Phages)

DNA viruses that infect bacteria

Phages and Pathology

Virulence Factors that cause cholera, dysentery, botulism, and food poisoning are carried on phage elements.



Novick, Richard, Plasmid (2003)

Objective

Characterize the abundance of bacterial virulence factors in phages

Data

Virulence Protein Databases

- VFDB
  Chen, Lihong, et al. Nucleic Acids Research (2005)
- PatricVF
  Wattam, AR, et al. Nucleic Acids Research (2017)

Virulence HMMs

- pFam
  Bateman, Alex, et al. Nucleic Acids Research (2004)
- pVOG
  Grazziotin, AL, et al. Nucleic Acids Research (2016)

Phage Protein Database

Methods

Sequence Annotation Methods

BLAST vs HMM

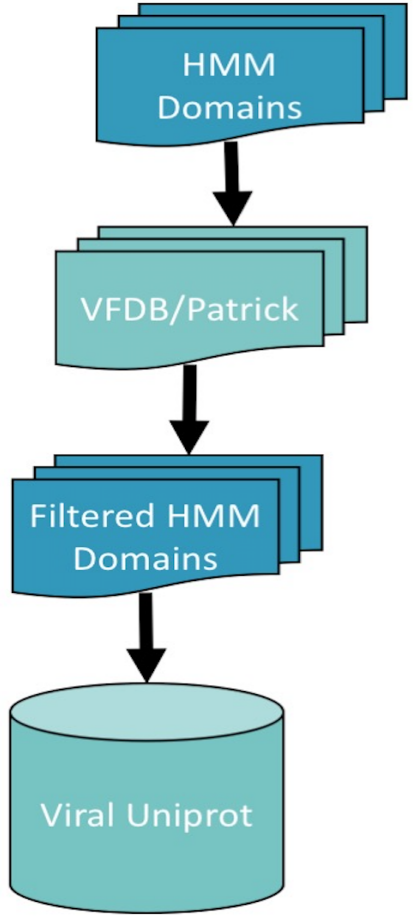
Normalizing By Gene Count

Hit Percentage = P
Hit Count = HC
Gene Count = GC

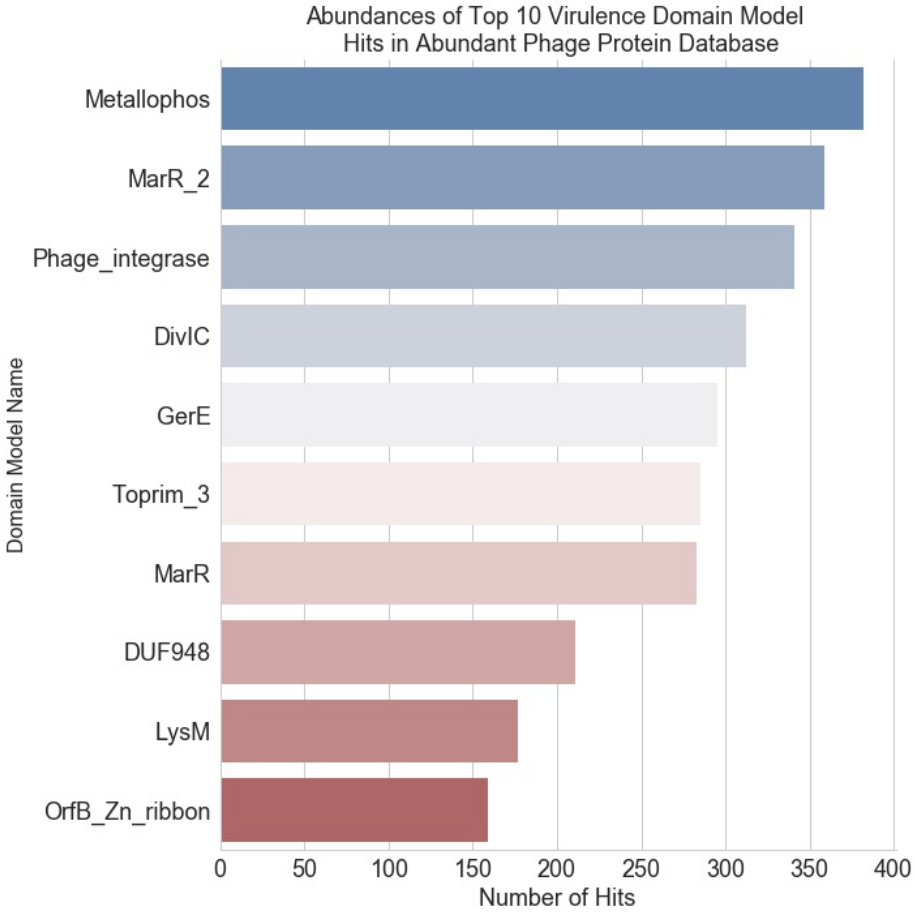
$$P = HC/GC$$

Filtering By Phage Abundance

Streptococcus phage:
Genera abundance greater than 30



HMM Hit Distribution



MarR

Domain involved in antibiotic resistance

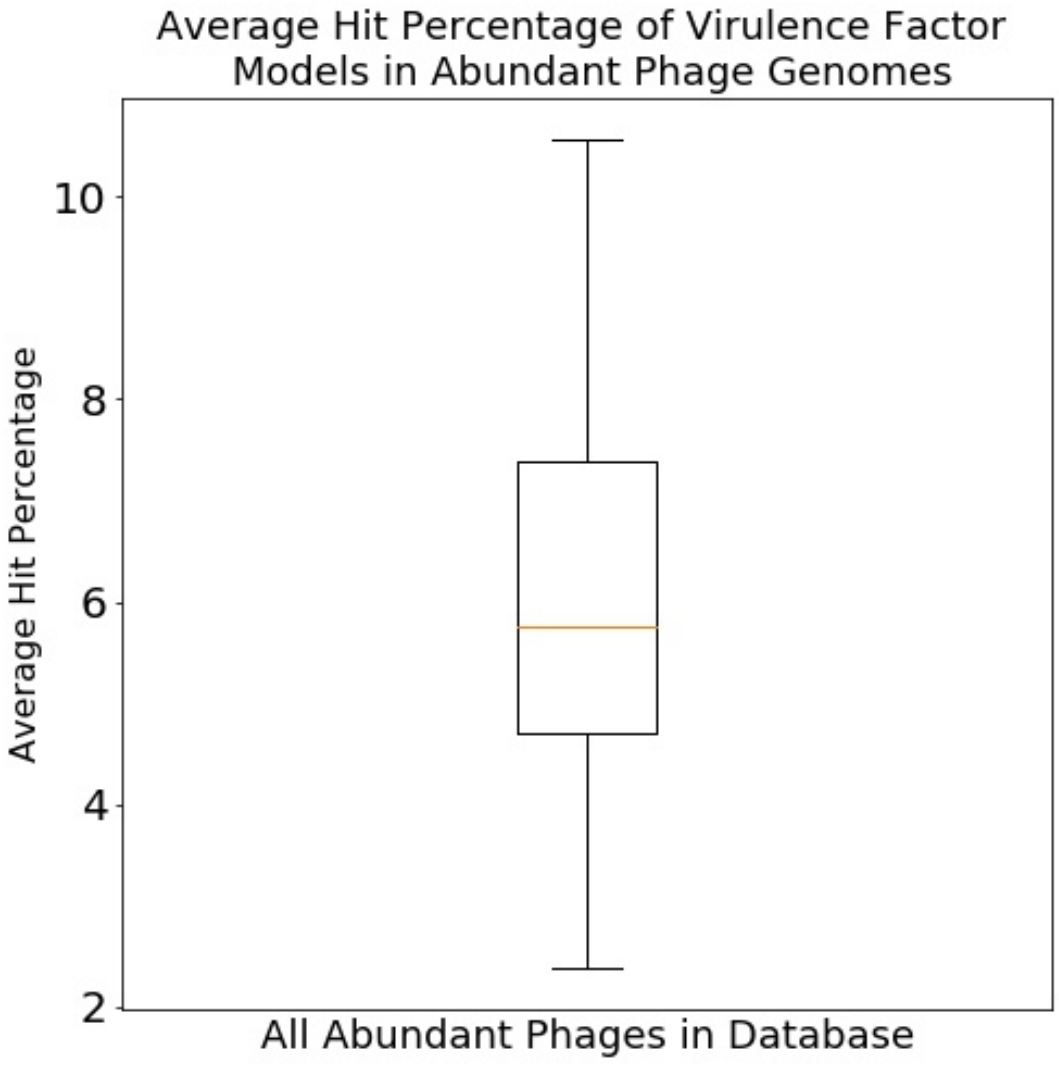
DivIC

Part of sporulation process

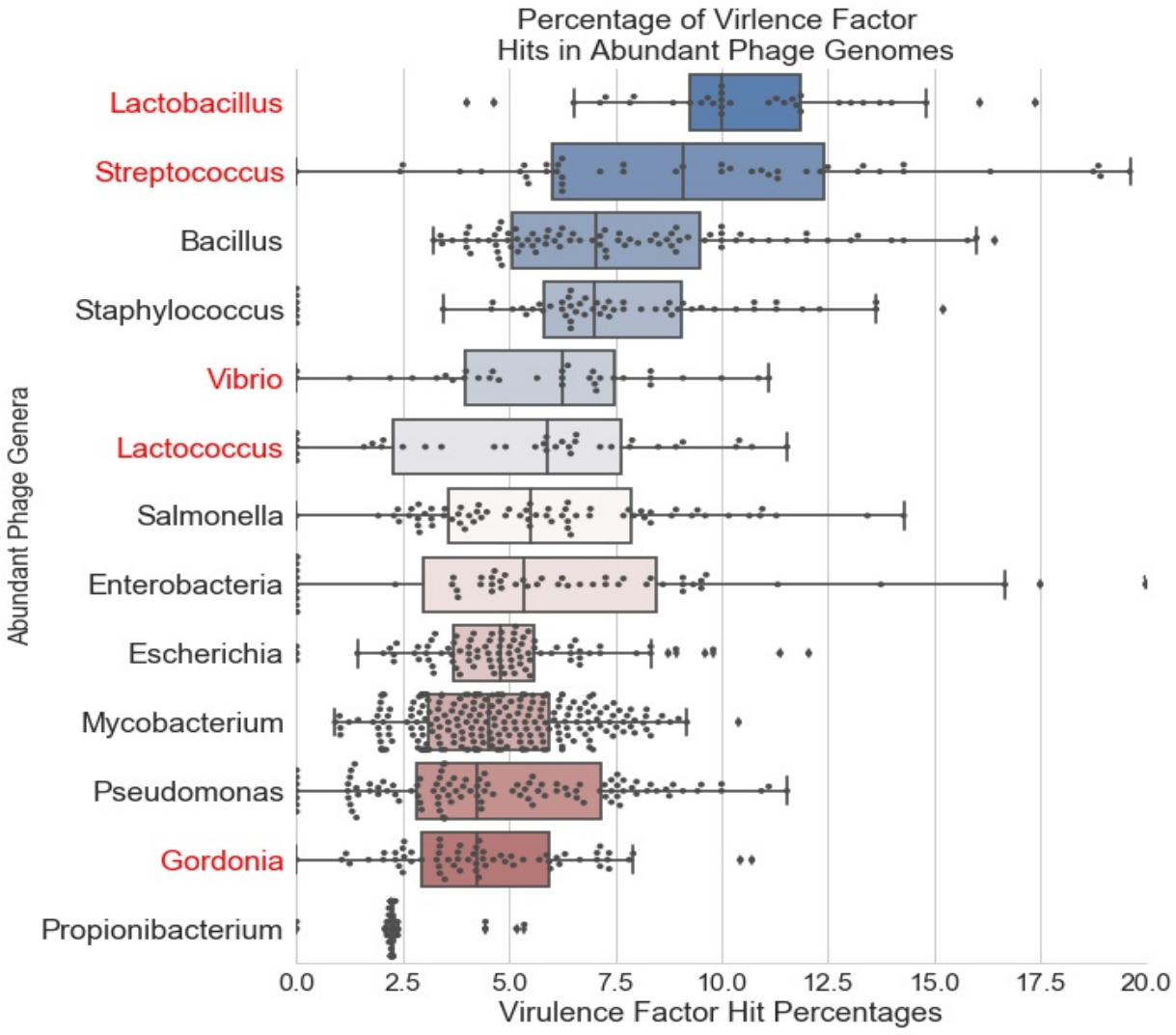
LysM

General peptidoglycan function

Distribution of Hit Percentage in All Phages



Abundant Phage Distributions by Genera Name



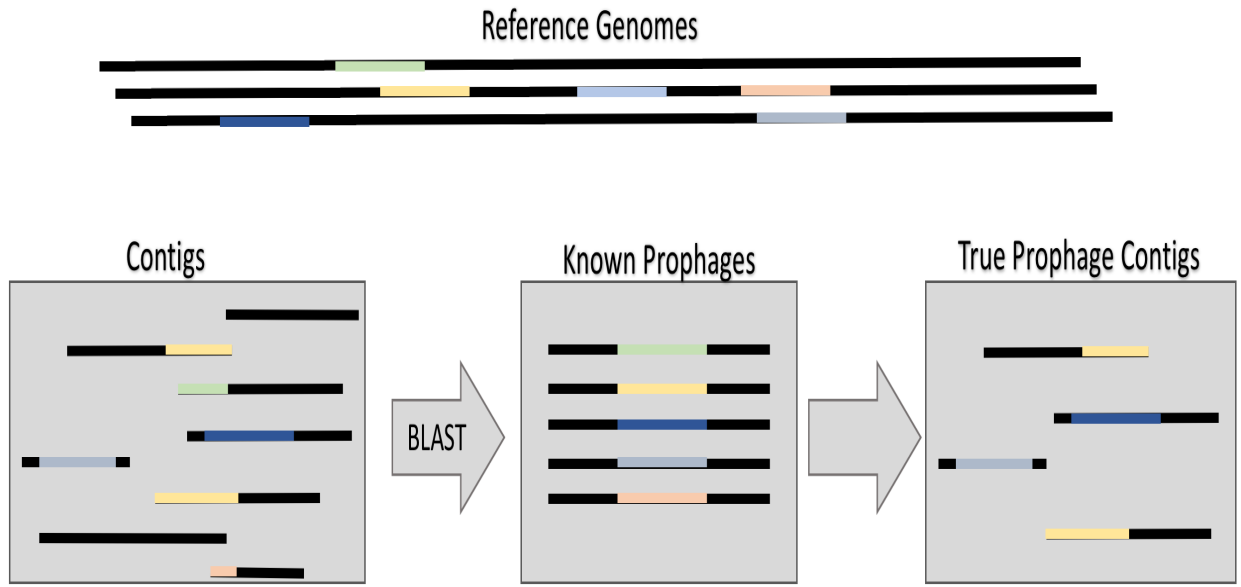
Future Directions

Magic-BLAST Streaming

Create a version of BUD for local metagenomic sequences

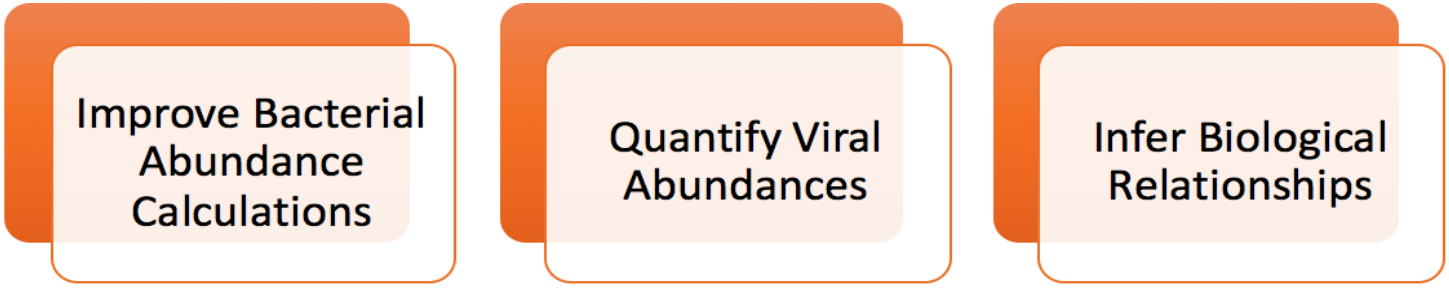
Testing Performance of BUD

Using the simulated dataset from previous study to compare the performance of identifying prophages by current tools against BUD



Contig Prediction

Concluding Remarks



Metagenomic Simulation Study

Effectively identifying viral elements improves bacterial abundance calculation

GRAB

Viral GRAB will contribute to a focus on phages specific to lung infections

Building Up Domains

Allows for the identification of prophages elements in metagenomics

Computational Bioscience Program

Elaine Epperson

Nabeeh Hasan

Josephina Hendrix

Michael Strong

Chris Miller

Cathy Lozupone

James Costello

Kirk Harris

Funding

NLM: 2 T15 LM 9451-11

Questions?

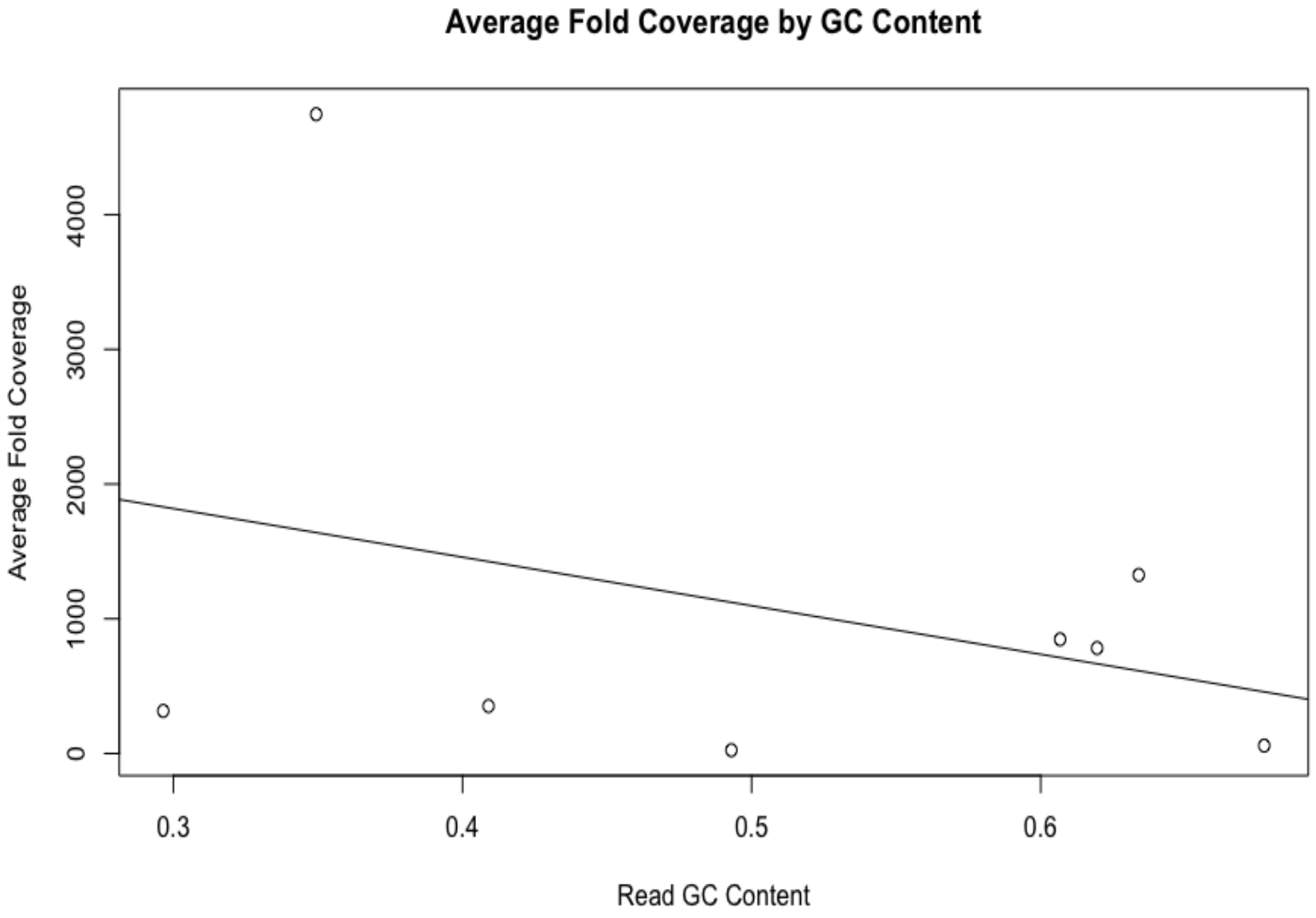
Cody Glickman

cody.glickman@ucdenver.edu

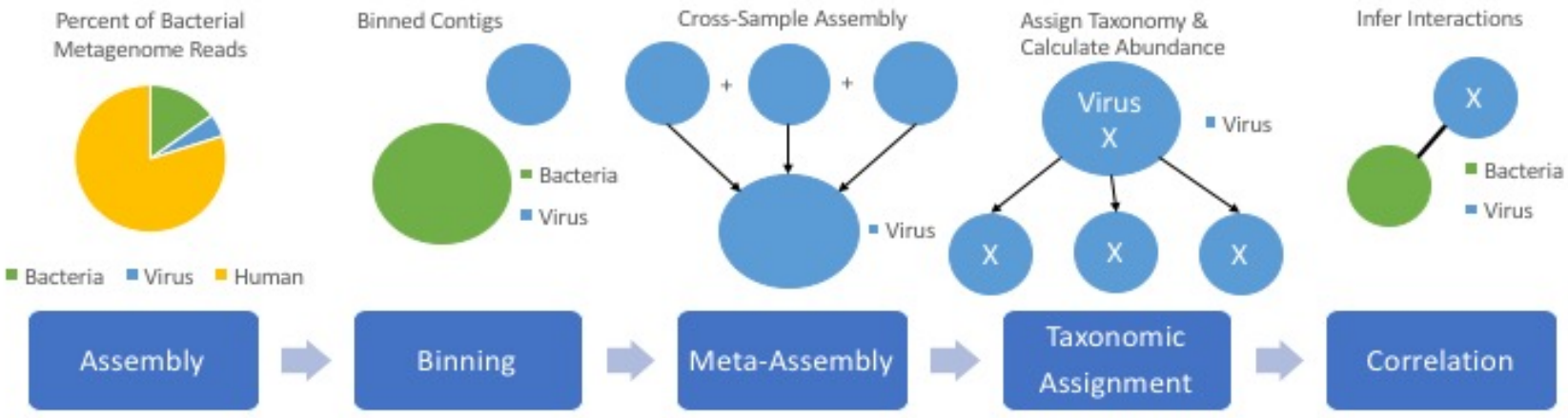
www.github.com/glickmac

www.codyglickman.com

Bias in Average Fold Coverage by GC



My Pipeline


Percent of Bacterial Metagenome Reads
Bacteria
Virus
Human
Assembly
Binned Contigs
Bacteria
Virus
Binning
Cross-Sample Assembly
Virus
Meta-Assembly
Assign Taxonomy & Calculate Abundance
Virus X
Virus
X
X
X
Taxonomic Assignment
Infer Interactions
X
Bacteria
Virus
Correlation

Tools Used in Study Continued

Simulation Tools

BBMAP - a suite of tools designed for sequencing data
Bushnell, B., JGI 2016

Taxonomic Identification

Kraken - A reference-free K-mer taxonomic identifier
Wood, Derrick E., and Steven L. Salzberg Genome 2014

Blastx - Referenced against a viral protein database
Camacho C., et al. BMC Bioinformatics 2008

Prophage Identification

Phaster - A popular prophage discovery web tool
Arndt, David, et al., Nucleic Acids Research 2016