

Introduction
ooooo

Metagenomic Simulation Study
oooooooooo

GRAB
oooooooo

Building Up Domains
oooooooo

Metagenomic Exploration: Evaluation and development of novel tools for viral and bacterial sequence analysis

Cody Glickman
CPBS Update Talk



Jan 29th, 2018

Introduction
ooooo

Metagenomic Simulation Study
oooooooooo

GRAB
oooooooo

Building Up Domains
oooooooo

Research Update

Asthma Environmental Microbiome

Building Up Domains

Clinical Features of Cystic Fibrosis and the Microbiome

Genomic Retrieval and Blast Database Creation

Hawaiian Soil Chemistry and Culture

Microbiome OTU Grouping Workflow

Metagenomic Simulation Study

Virome Statistics Pipeline

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

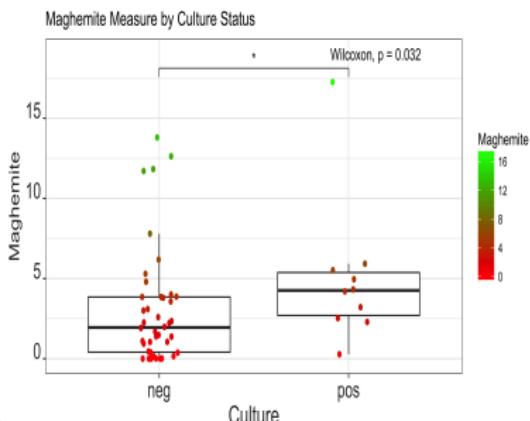
Strollo SE, et al. Ann Am Thorac Soc. 2015

Adjemian J, et al. Am J Respir Crit Care Med. 2012

Understanding Why NTM Develops

Hawaiian Soil Project

Identifying important soil characteristics for NTM soil culture



Pulmonary NTM

- 90% of NTM cultures are from respiratory samples
- The lung has the lowest abundance of bacteriophages among human niches

O'Brien, R., et al. American Review of Respiratory Disease 1987

Aziz R., et al. Frontiers in microbiology 2015

Of "Viral" Importance

Bacteriophages (Phages)

Phages are DNA viruses that infect prokaryotes

Bacteriophage Adherence to Mucus (BAM)

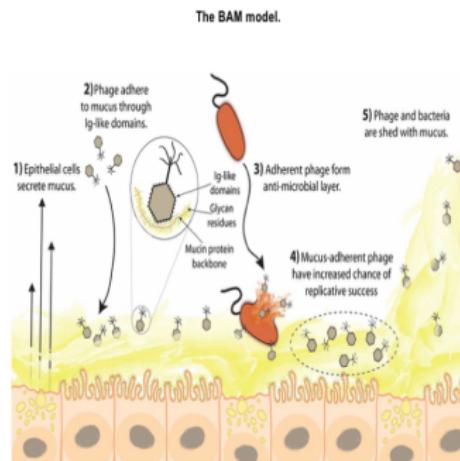
- Phages act as an innate immune system in mucosal tissues
- Prior studies identified Ig-like motifs in induced phages from *Pseudomonas* cultures

Phages in the Lungs

Does phage abundance in the lungs affect bacterial biofilm formation?

Barr, Jeremy, et al., PNAS 2013

Tariq, Mohammad, et al., Frontiers in Microbiology 2015



Jeremy J. Barr et al. PNAS 2013;110:10771-10776

©2013 by National Academy of Sciences

PNAS

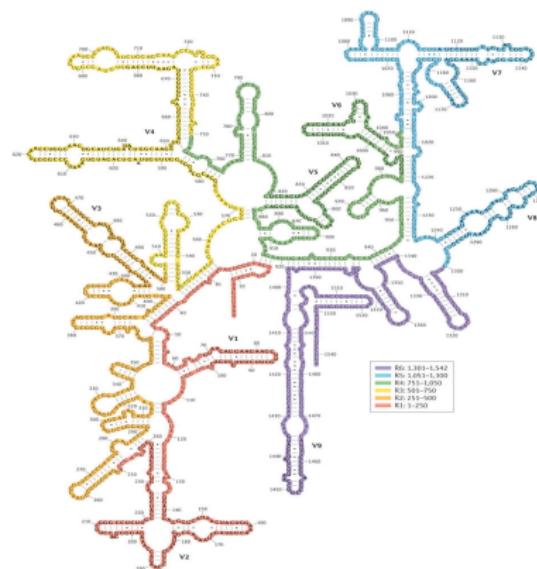
Molecular Methods to Study Phages

Difficulties of phage study

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

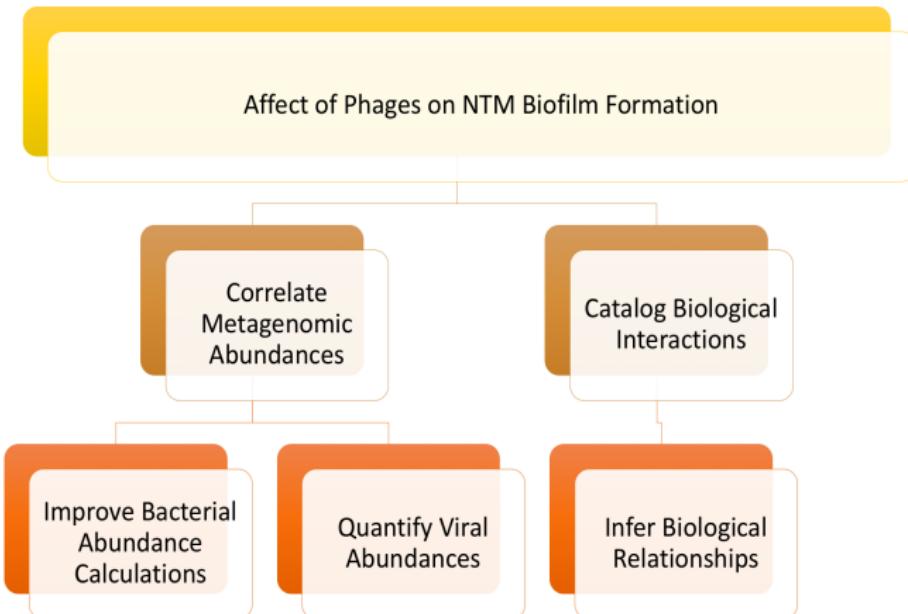
Phage Isolation Methods

- Biological filtration
- In silico methods



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

Objective



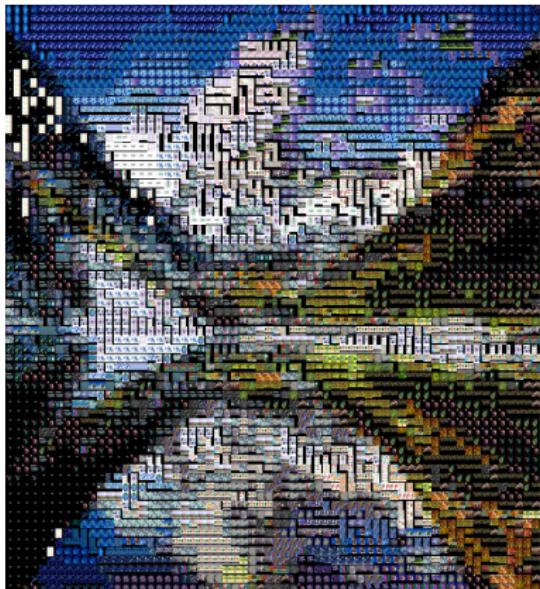
Metagenomics

What is Metagenomics?

Unbiased study of all genetic material in a sample

Importance of Metagenomics

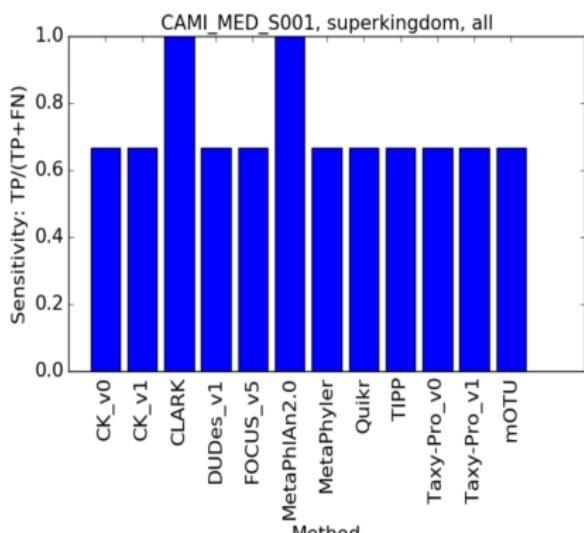
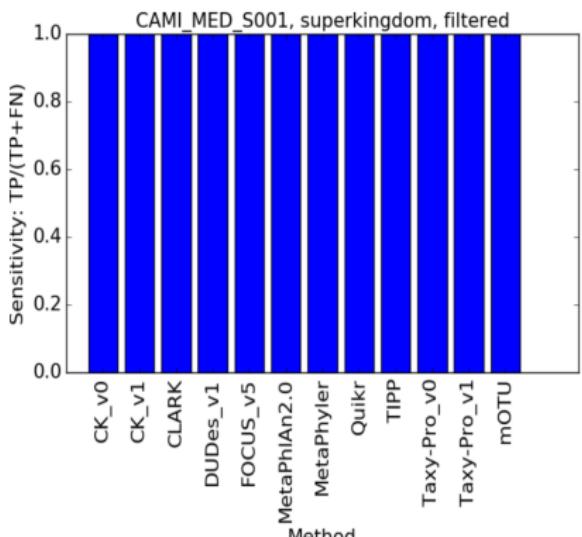
- Functional capabilities of a sample
- Species level distinctions
- Due to lack of a universal gene marker, phages are studied by metagenomics



Metagenomics Gold Standard

Critical Assessment of Metagenome Interpretation (CAMI)

- Comprehensive simulation study of tools from all levels of analysis (binners, assemblers, taxonomic profilers)
- Viral and plasmids affected the performance of taxonomic profilers and abundance calculations



Viral Filtration Simulation Study

Study Design

30 simulated mixed metagenomes are used to compare the viral contiguous sequence (contigs) identification performance of multiple tools

Sequencing Depth of Experiment

Each metagenome is comprised of 10 million reads

Complexity of Metagenomes

8 bacteria and 8 phages comprise the low complexity samples in each metagenome

Genomes in Simulation

Virus - 0.12 Mb

- Bacillus phage Pony
- Caulobacter phage CcrColossus
- Mycobacterium phage Bxb1
- Mycobacterium phage Che9d
- Mycobacterium phage TM4
- Pseudomonas phage vB-PaeM-C2-10-Ab1
- Staphylococcus phage CNPH82
- uncultured phage crAssphage

Bacteria - 4.72 Mb

- Bacillus subtilis subs. subtilis 168
- Clostridium acetobutylicum ATCC 824
- Clostridium perfringes str. 13
- Lactococcus lactis subsp. lactis II1403
- Pseudomonas aeruginosa LESB58
- Staphylococcus aureus subsp. aureus N315
- Streptococcus pyogenes M1 476
- Xylella fastidiosa 9a5c

Introduction
ooooo

Metagenomic Simulation Study
oooo●oooo

GRAB
oooooooo

Building Up Domains
oooooooo

Tools Used in Study

Assembler

MEGAHIT - Effective at assembling viromes

Roux, S., et al. PeerJ 2017

Viral Contig Identification Methods

VirFinder - Viral contig K-mer identification model

Ren, Jie, et al. Microbiome 2017

Blastx - Filtering against a viral protein database

Camacho C., et al. BMC Bioinformatics 2008

VirSorter - Hybrid HMM and gene marker database

Roux, S., et al. PeerJ 2015

vHMM - Iterative HMM trained on viral genes

Paez-Espino, D., et al. Nature 2016

Performance Measurements

True Viral Contigs

True viral contigs are defined by BLAST hits (E-value 10^{-5}) against a custom database of the reference phages and bacterial prophage elements.

Term Definitions

TP = True Positive FP = False Positive FN = False Negative

Performance Metrics

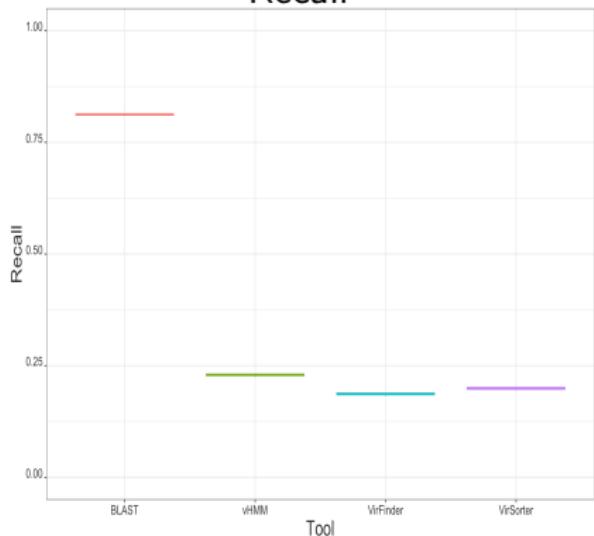
- Recall = $TP / (TP + FN)$
- Precision = $TP / (TP + FP)$

Recall and Precision

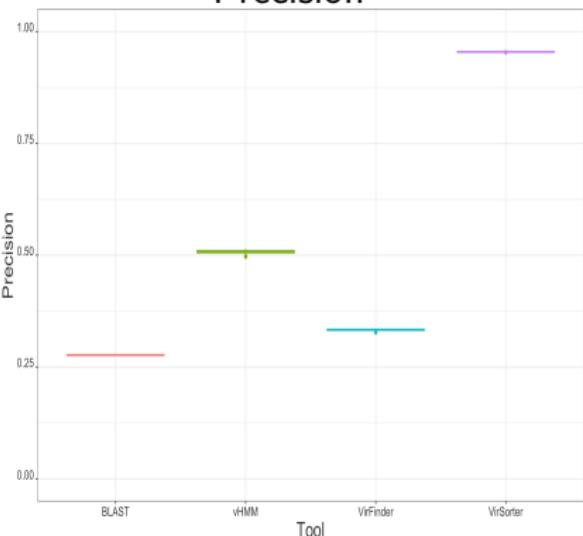
Order of Tools
from left to right

BLAST, vHMM, VirFinder, VirSorter

Recall



Precision



Conclusions

Performance

The variance of tool performance suggests that no single tool is optimal for viral filtration

Tool Parameter Optimization

Used default tool parameters for all tools

- Lenient BLASTX filtering resulting in false positives
- vHMM model is the second iteration

Introduction
ooooo

Metagenomic Simulation Study
oooooooo●

GRAB
oooooooo

Building Up Domains
oooooooo

Future Directions

Expansion of Tools

Inclusion of binners MetaBat and MetaWatt 3.5

Expanded Dataset

Filter viral elements from CAMI data

Combinations of Tools

Explore how the tools perform in combinations

Creating Custom BLAST Databases

Demand for Custom Databases

The Strong Lab uses custom databases to identify mycobacterium elements in metagenomics

Current Tools

- `makeblastdb` from Command-line BLAST allows users to create custom BLAST databases
- `makeblastdb` requires local sequences to create usable database

Batch Sequence Retrieval

Current Methods

- NCBI Webserver
- ESearch / Efetch function
- biomartr R package

Test Case

Retrieve protein sequences of three *Mycobacterium* subspecies

- *Mycobacterium avium paratuberculosis*
- *Mycobacterium abscessus massiliense*
- *Mycobacterium abscessus bolletii*

NCBI Webserver

Batch Download Process

- Query Assembly: "Mycobacterium abscessus **subsp.** massiliense"[Organism] OR "Mycobacterium abscessus **subsp.** bolletii"[Organism] OR "Mycobacterium avium **subsp.** paratuberculosis"[Organism]
- Select Complete Genome Tab
- Click "Download Assemblies" button and select Protein FASTA from File Type drop down

Difficulties

- Query string must be specific
- Length of query string can become unwieldy
- Sparse information on download procedure and examples

Efetch and biomartr

Efetch Retrieval Process

Efetch can be utilized via Python or via command-line

```
handle = Entrez.efetch(db="nuccore", id= UID_List ,  
rettype = "fasta_cds_aa")
```

biomartr Sequence Retrieval

```
library(biomartr)  
meta.retrieval(kingdom = "bacteria", group =  
"Actinobacteria", db = "genbank", type = "proteome")
```

Difficulties

- Entrez.efetch function requires *a priori* knowledge of Entrez UID
- biomartr only able to bulk download phylum level
- Programming skills important [not required for Efetch]

Genomic Retrieval and Blast Database Creation (GRAB)

A Batch Retrieval System for Biologists

Well documented command line tool and web interface (in development)

GRABs Sequences by Taxonomy

- GRAB requires the name of organisms and the taxonomic level
- GRAB retrieves genomic, coding sequences, or protein sequences

GRAB Sequence Retrieval

```
python GRAB.py -m protein -q  
paratuberculosis,massiliense,bolletii -l subspecies
```

Introduction
ooooo

Metagenomic Simulation Study
ooooooooo

GRAB
ooooo●oo

Building Up Domains
ooooooo

Genomic Retrieval and Blast Database Creation

GRAB Pipeline

Introduction
ooooo

Metagenomic Simulation Study
ooooooooo

GRAB
oooooo●○

Building Up Domains
oooooooo

Demand for GRAB



Nine questions related to bulk download of protein or genomic sequences (Over 58K views)



Four question related to bulk download of protein or genomic sequences (Over 3.5K views)

Future Directions

Webserver

Shiny web application in development

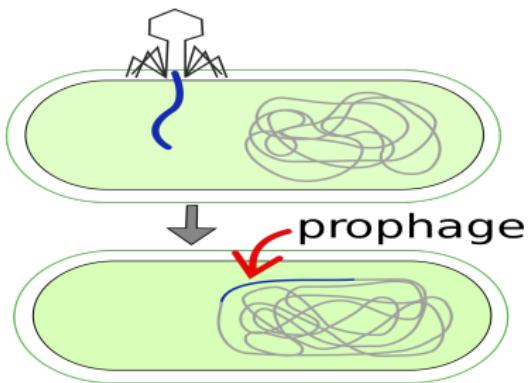
Viral GRAB

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

Endogenous Viral Elements (Prophages)

Lysogenic Life Cycle

Viruses can integrate into host for an extended period of time



Importance of Prophages

Prophages can confer advantages to host improving survival

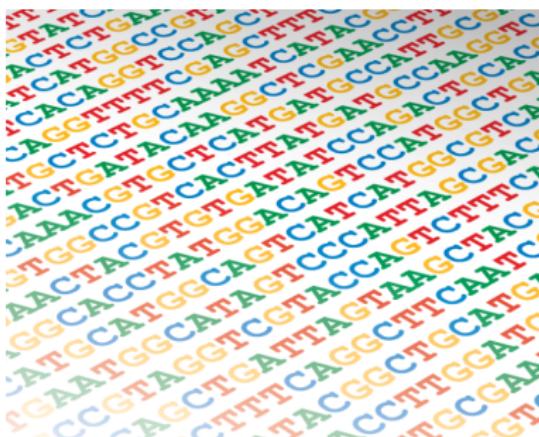
Prophages are important to the emergence of pathogenic bacteria

Canchaya C., et al. *Curr Opin Microbiol* 2003
Wagner PL. & Waldor MK. *Infect Immune* 2002

Finding Prophages

Prophage Discovery Problem

Same difficulties as gene prediction: finding signal in data



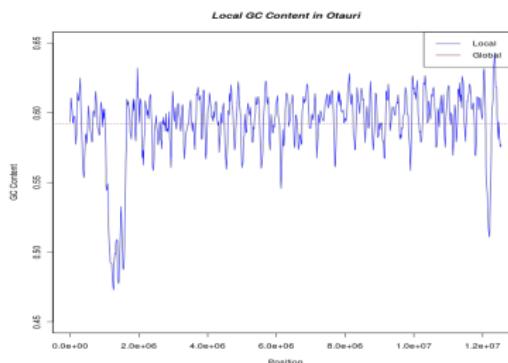
Prophage Discovery Tools

Number of Tools

10 tools listed at Omic Tools for prophage discovery

Methods Used

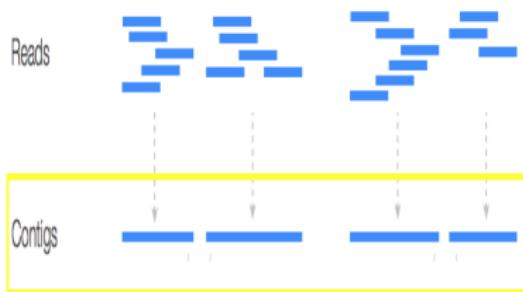
- Sequence similarity
- Hidden Markov models
- Transcription direction
- Protein length
- Sliding window GC content
- Phage specific kmer



Prophage Discovery Methods

Top Down Methods

All prophage discovery methods find prophages within contiguous sequences or genomes



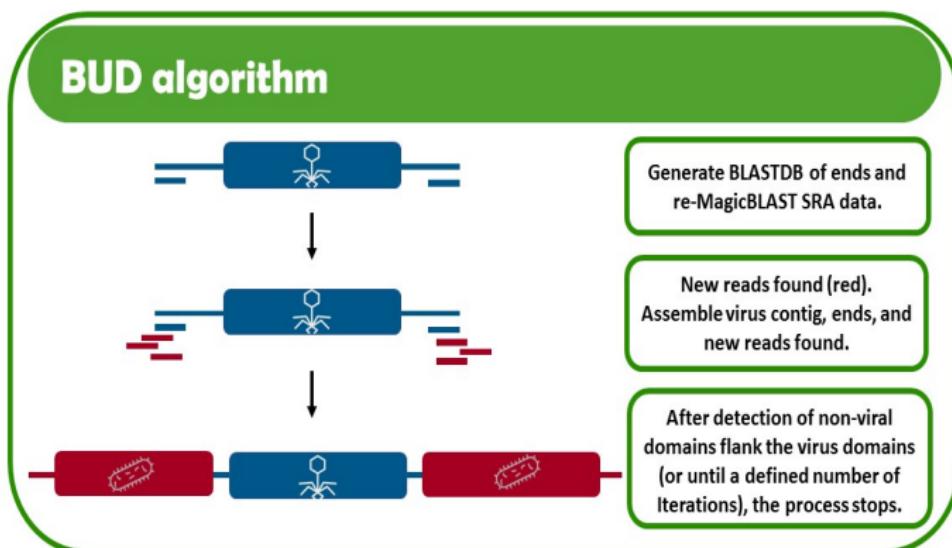
Potential Prophage Tool Pitfalls

- Metagenomics produces short contigs that are discarded
- In metagenomics, prophage hosts may not be identified

Building Up Domains (BUD) Algorithm

Initialization

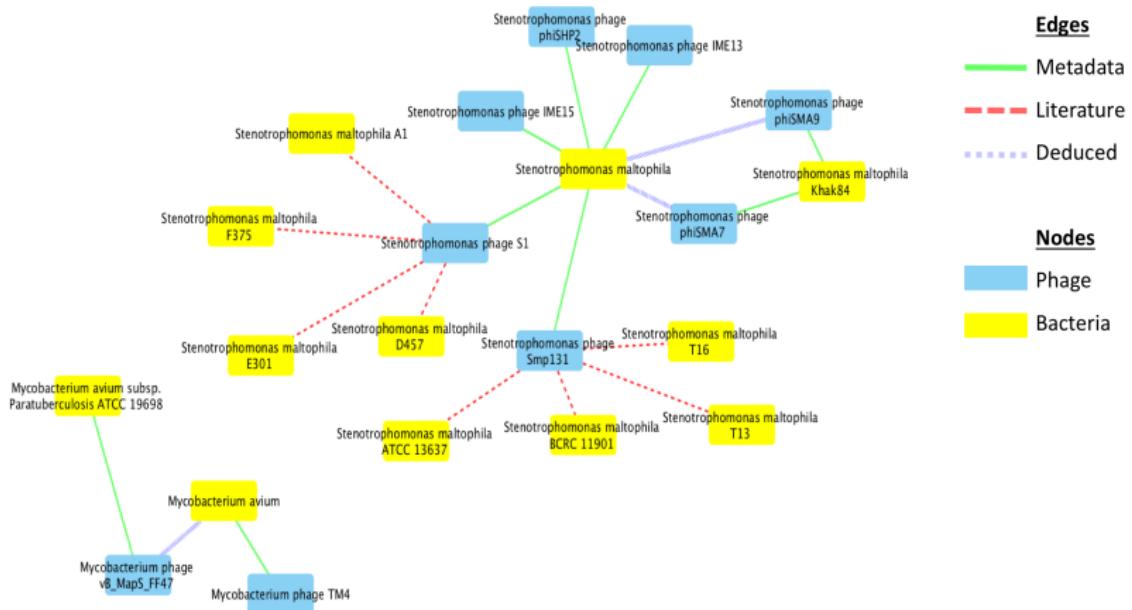
- Metagenomic reads are filtered by BLAST against Viral RefSeq
- Remaining reads are assembled into contigs



Potential Uses of BUD

Expanding Known Phage Host Range

BUD has the potential to identify novel hosts for prophages



Current Implementations of BUD

ViruSpy

- Originally written for NCBI Hackathon
- BUD Algorithm written in Perl and BASH
- Utilized Magic-BLAST for streaming of reads

EndoVir

NCBI Collaborators Jan Buchman and Ben Busby

- Written in Python
- Implementation of BUD with Magic-BLAST

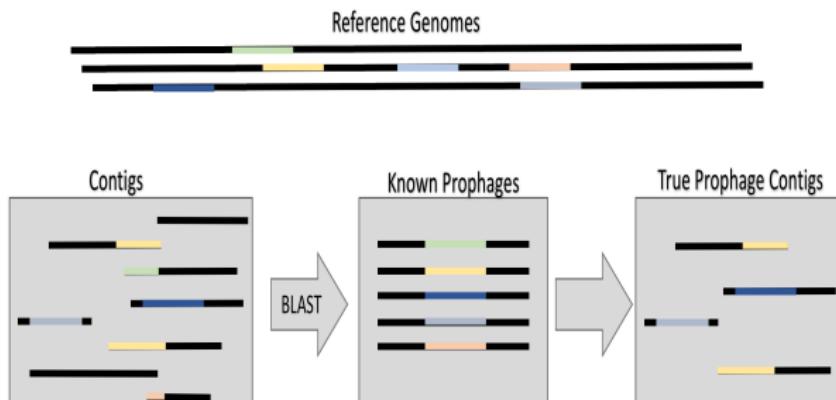
Future Directions

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



Concluding Remarks

Improve Bacterial
Abundance
Calculations

Quantify Viral
Abundances

Infer Biological
Relationships

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

Chris Miller

Nabeeh Hasan

Cathy Lozupone

Josephina Hendrix

James Costello

Michael Strong

Kirk Harris

Funding

NLM: 2 T15 LM 9451-11



Introduction
ooooo

Metagenomic Simulation Study
oooooooooo

GRAB
oooooooo

Building Up Domains
oooooooo

Questions?

Cody Glickman



cody.glickman@ucdenver.edu

www.github.com/glickmac

www.codyglickman.com

Introduction
ooooo

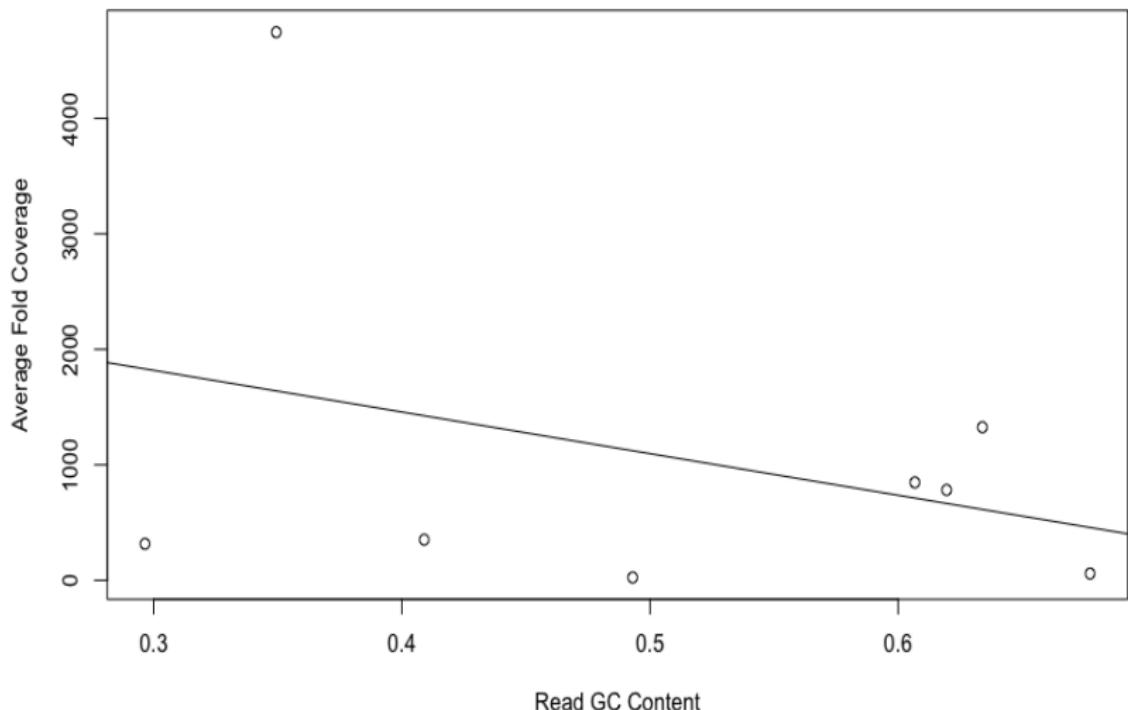
Metagenomic Simulation Study
oooooooooo

GRAB
oooooooo

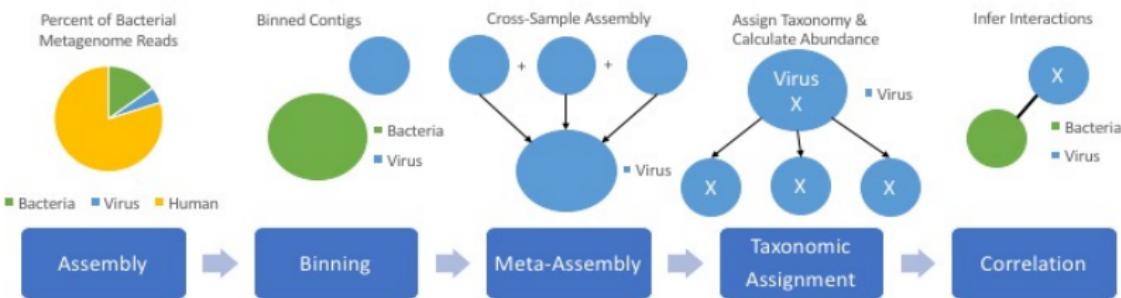
Building Up Domains
oooooooo

Bias in Average Fold Coverage by GC

Average Fold Coverage by GC Content



My Pipeline



Introduction
ooooo

Metagenomic Simulation Study
ooooooooo

GRAB
oooooooo

Building Up Domains
oooooooo

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