Hodgepodge Metagenomics: A collection of novel tools for viral and bacterial sequences

Cody Glickman
CPBS Update Talk

STRONG LAB

Jan 29th, 2018

Table of Contents

Non-Tuberculosis 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

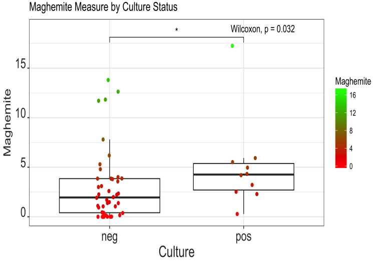
- Immunocompromised individuals
- Patients with lung damage or malfunction
- Residing in warm costal areas especially Hawaii

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 DNA viruses in the human niche

O'Brien, R., et al. American Review of Respiratory Disease 1987
Aziz R., et al. Frontiers in microbiology 2015

Of "Viral" Importance

Bacteriophages aka Phages

Phages are DNA viruses that infect prokaryotes

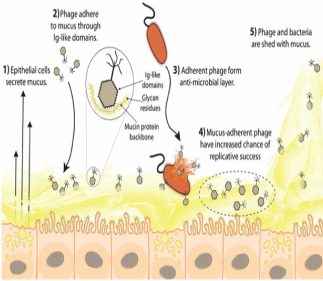
Bacteriophage Adherence to Mucus (BAM)

- 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

The abundance of phages is significantly lower in the lungs

The BAM model.


1) Epithelial cells secrete mucus.
2) Phage adhere to mucus through Ig-like domains.
Ig-like domains
Glycan residues
Mucin protein backbone
3) Adherent phage form anti-microbial layer.
4) Mucus-adherent phage have increased chance of replicative success
5) Phage and bacteria are shed with mucus.

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

PNAS

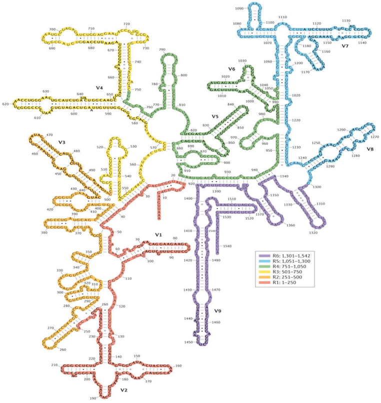
Molecular Methods to Study Phages

Difficulties of phage study

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

Phage Isolation Methods

- Biological filtration
- In Silico Methods



Nature Reviews | Microbiology

Objective

Develop new tools and incorporate them into pipelines to identify and quantify bacteriophage elements in shotgun metagenomic sequences.

Secondary Goal

Identify relationships between bacteria and phages using abundance quantification across multiple studies.

Metagenomics

What is Metagenomics?

The study of genetic material from environment or clinical samples

Importance of Metagenomics

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

Include Photo Mosaic Here

Metagenomics Gold Standard

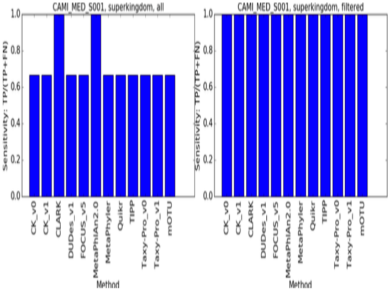
Critical Assessment of Metagenome Interpretation (CAMI)

- 1300 newly sequenced organisms collected in simulated dataset
- 25 programs and 36 biobox implementations (binners, assemblers, taxonomic profilers)

Pyrite Standard

Viral elements worsened abundance estimates in taxonomic profilers

Sczyrba, A., et al. Nature Methods 2017

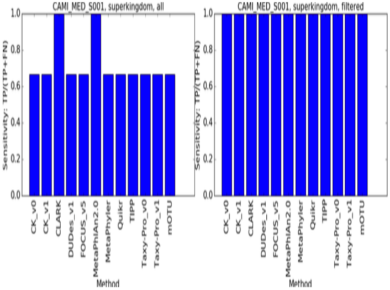


Supplementary Figure 41. Sensitivity metric at the superkingdom rank for each profiler on the low complexity, unfiltered sample (left) and filtered sample (right).

Improving Taxonomic Profiling

CAMI and Viruses

Filtering viruses improves abundance profile estimates for bacteria



Supplementary Figure 41. Sensitivity metric at the superkingdom rank for each profiler on the low complexity, unfiltered sample (left) and filtered sample (right).

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 Il1403
- Pseudomonas aeruginosa LESB58
- Staphylococcus aureus subsp. aureus N315
- Streptococcus pyogenes M1 476
- Xylella fastidiosa 9a5c

Tools Used in Study

The tools used in this study are selected based on recent publications

Assembler

MEGAHIT - Effective at assembling viromes
Roux, Simon, et al. PeerJ 2017

Filtration 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

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

Performance Measurements

True Viral Contigs

True viral contigs are defined by BLAST hits (E-value 10ˆ05) against a custom database of reference phages and bacterial prophage elements.

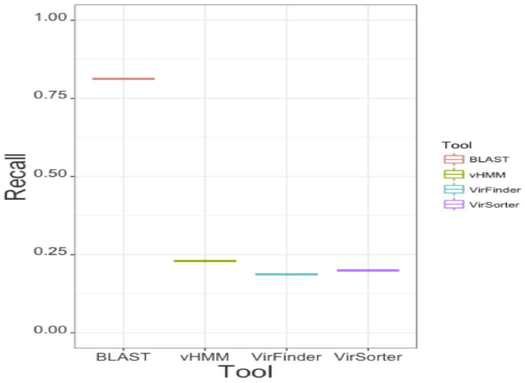
Term Definitions

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

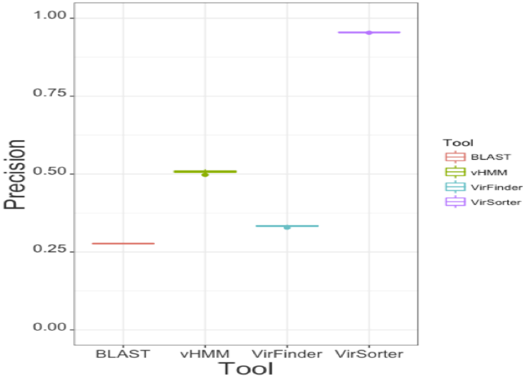
Performance Metrics

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)
- F1 = (2*TP) / (2*TP + FP + FN)

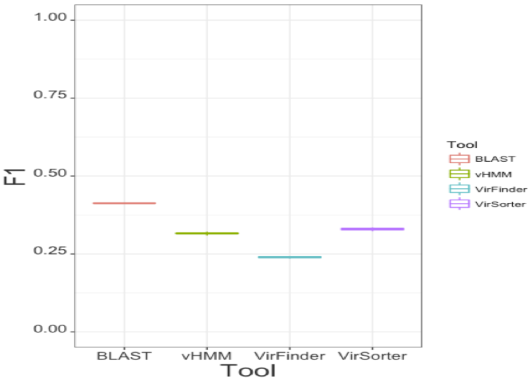
Recall


1.00
0.75
0.50
0.25
0.00
Recall
BLAST
vHMM
VirFinder
VirSorter
Tool
Tool
BLAST
vHMM
VirFinder
VirSorter

Precision


Precision
1.00
0.75
0.50
0.25
0.00
BLAST
vHMM
VirFinder
VirSorter
Tool
Tool
BLAST
vHMM
VirFinder
VirSorter

F1


1.00
0.75
0.50
0.25
0.00
F1
BLAST
vHMM
VirFinder
VirSorter
Tool
Tool
BLAST
vHMM
VirFinder
VirSorter

Conclusions and Future Directions

Performance

The variance of performances suggests that no one tool is optimal for viral filtration

Expansion of Tools

Inclusion of binners MetaBat and MetaWatt 3.5

Expanded Dataset

Filter viral elements from CAMI data

Tool Parameter Optimization

VirFinder is currently trained against only phages

Custom BLAST Databases

Command-line BLAST

Released in 2008 to allow users to run BLAST on local machines

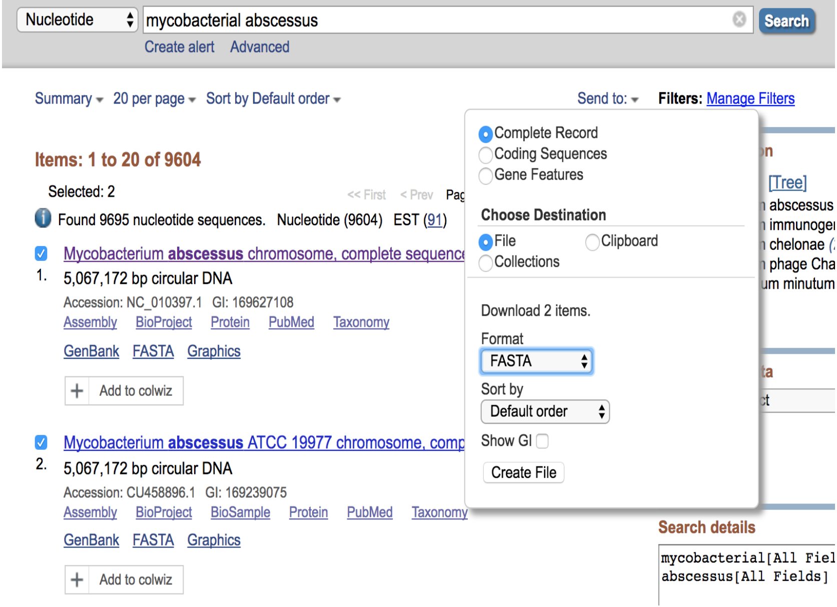
makeblastdb Function

Allows user to create custom BLAST databases from local sequences

Sequence Batch Retrieval

NCBI Webserver, ESearch function, biomart R package

Current Batch Retrieval



Nucleotide ⬍ mycobacterial abscessus ⊗ Search

Create alert Advanced

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾ **Filters:** Manage Filters

Items: 1 to 20 of 9604

Selected: 2

<< First < Prev Pag

ⓘ Found 9695 nucleotide sequences. Nucleotide (9604) EST (91)

☑ Mycobacterium **abscessus** chromosome, complete sequence

1. 5,067,172 bp circular DNA

Accession: NC_010397.1 GI: 169627108

Assembly BioProject Protein PubMed Taxonomy

GenBank FASTA Graphics

+ Add to colwiz

☑ Mycobacterium **abscessus** ATCC 19977 chromosome, comp

2. 5,067,172 bp circular DNA

Accession: CU458896.1 GI: 169239075

Assembly BioProject BioSample Protein PubMed Taxonomy

GenBank FASTA Graphics

+ Add to colwiz

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections

Download 2 items.

Format
FASTA ⬍

Sort by
Default order ⬍

Show GI □

Create File

n

[Tree]
abscessus
immunoge
chelonae (
phage Cha
um minutum

ta

ct

Search details

```
mycobacterial[All Fiel
abscessus[All Fields]
```

Obstacles to Current Systems

Taxonomic Querying

Querying multiple bacteria by taxonomy requires long command

```
"Mycobacterium abscessus"[Organism] OR "Mycobacterium avium"[Organism]
```

A Priori Knowledge

ESearch requires knowledge of accession numbers to query

Required Programming Knowledge

The biomart R Package requires knowledge of IDE and data manipulation

Genomic Retrieval and Blast Database Creation (GRAB)

A Batch Retrieval System for Biologists

Well-Documented Command Line and Web interface

Allows for Taxonomic Querying

Acknowledgements

STRONG LAB

Elaine Epperson

Michael Strong

National Jewish Health

HongWei Chu

Davidson Lab

Rebecca Davidson

Computational Bioscience Program

computational bioscience program

school of medicine at the university of colorado denver

Chris Miller

Cathy Lozupone

James Costello

Kirk Harris

Funding

NLM: 5 T15 LM 9451-09

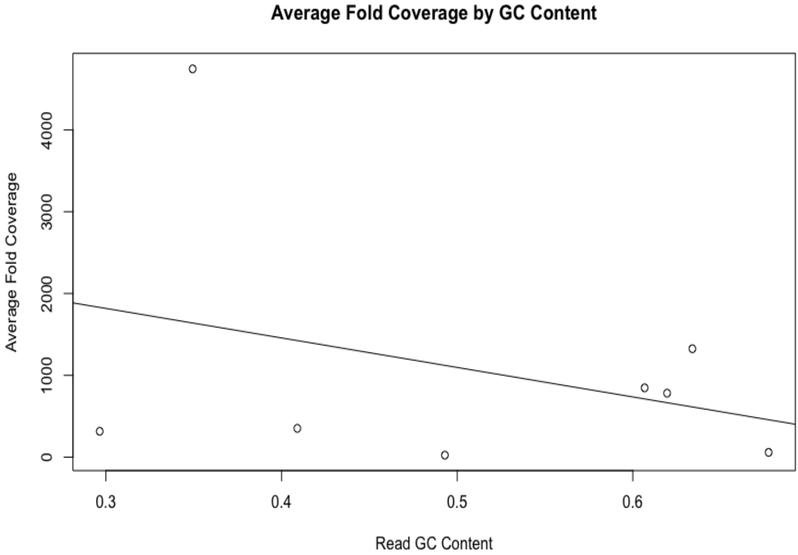
Questions?

Cody Glickman

STRONG LAB

cody.glickman@ucdenver.edu
www.github.com/glickmac
www.codyglickman.com

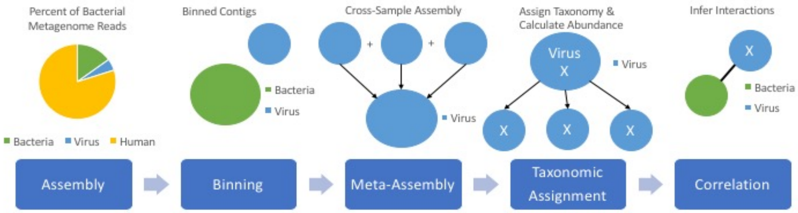
Bias in Average Fold Coverage by GC



References

Barr, Jeremy, et al., PNAS 2013
Tariq, Mohammad, et al., Frontiers in Microbiology 2015

My Pipeline


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