

TBD

Cody Glickman
Journal Club



Nov 16, 2017

Table of Contents

Introduction

Methods

Simulation Study

Sensitivity Study

NTM Prevalence

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

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

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

Population At Higher Risk of Developing NTM

Immunocompromised Individuals

HIV / AIDs

Individuals with Lung Damage

Cystic Fibrosis (CF)

Bronchiectasis

Location, Location, Location

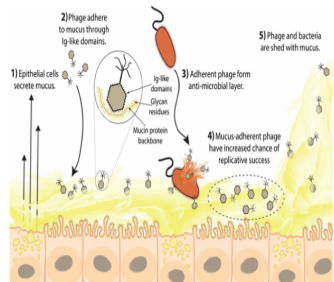
Warm costal areas

Infections and the Virome

Bacteriophage Adherence to Mucus (BAM)

- Viruses in mucosal outnumber bacteria 40 to 1
- Viruses act as an innate immune system in the mucosal
- Previous studies identified Ig-like motifs in induced phages from *Pseudomonas* cultures

The BAM model.



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

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

Molecular Methods to Study Virome

- Filtration + DNase
- Dithiothreitol + Filtration + DNase
- Filtration + DNase + CsCL Centrifugation
- In Silico Methods

The first three methods are subject to inherent sampling biases

Kleiner, M., et al BMC genomics 2015

Hypothesis

Implementation of a filtration process will improve performance of taxonomic identification of viral elements in bacterial metagenomics

Study Design

This study establishes the feasibility for the filtration and novel viral identification pipeline in development.

Simulation Study

A simulated mixed metagenome is used to compare the viral taxonomic identification performance

Sensitivity Study

A real longitudinal metagenomic dataset is spiked with a rare virus to measure sensitivity of taxonomic assignments.

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

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

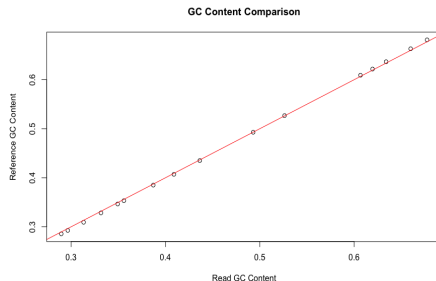
- Bacillus subtilis subs. subtilis 168
- Clostridium acetobutylicum ATCC 824
- Clostridium perfringes str. 13
- Lactococcus lactis subsp. lactis II1403
- Pseudomonas aeruginosa LESB58
- Staphylococcus aureus subsp. aureus N315
- Streptococcus pyogenes M1 476
- Xylella fastidiosa 9a5c

Simulation Details

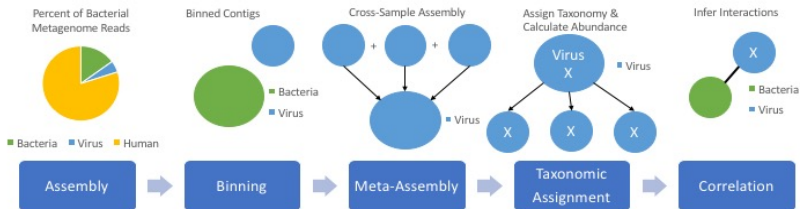
Library Size 10 Million Reads

Insert Size 150 BPs non-paired

Error Rate No errors introduced



My Pipeline



Performance Measurements

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

Precision

$$P = \frac{TP}{TP + FP}$$

Recall

$$R = \frac{TP}{TP + FN}$$

F1 Score

$$F1 = \frac{2(TP)}{2(TP) + FP + FN}$$

Results

Performance of methods identifying viral elements in simulated metagenome

	Precision	Recall	F1 Score
Raw Reads	0.0593	1	0.1119
Full Assembly	0.3478	1	0.5161
Filter Pipeline	0.4615	0.75	0.5714
Blastx Filter	0.4444	0.5	0.4706

Table: The F1 performance of the filter pipeline exceeds all other methods. The filtration method trades recall for overall performance.

Troubleshooting

Prophages

The bacterial genomes selected all contain prophage elements

Casjens, Sherwood. Molecular microbiology 2003

Prophage Discovery

The web-tool Phaster collected prophage prediction taxonomy on genomes used in simulation

No overlap of FP viruses and prophages predicted (Performed on Assembly and Filtered only)

Data

A longitudinal survey of the Cystic Fibrosis airway of a single patient

Number of Samples 36 Samples

Avg Library Size 33.3 Mb per Sample

Insert Size 300 BPs non-paired (454
pyrosequencing)

Read Composition Samples pre-filtered human samples
using Deconseq

Schmieder, Robert, and Robert Edwards. PLoS one 2011

Synthetic Spike-In

To test sensitivity of pipeline added a rare virus to real dataset

Zaire ebolavirus

18.96 Kb genome size

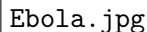
Generated 2000 reads using BBMap

Synthetic Assembly

Generated a single contig 18.93Kb

Distributed Reads

Incorporated 56 random ebola reads into samples

A rectangular box containing the text "Ebola.jpg".

Results

The results are based on presence absence in kraken taxonomic identities

Combined Sample Assembly

Identifies Zaire ebolavirus

Significant Viral Contigs

Absent

Viral Reads and Assembled

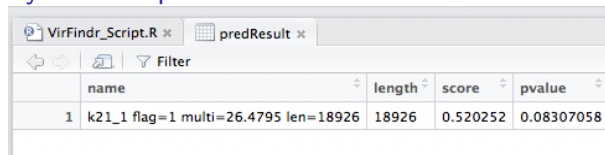
Absent

Discussion and Future Directions

Synthetic Metagenome

- The taxonomic identification performance of the filtration model exceeds that of both the raw and assembled reads.
- Increasing the complexity of the simulation by both adding mutations and increasing the number of genomes is planned for this week.

Synthetic Spike-In



	name	length	score	pvalue
1	k21_1 flag=1 multi=26.4795 len=18926	18926	0.520252	0.08307058

Acknowledge.jpg

Questions?

Cody Glickman



cody.glickman@ucdenver.edu

www.github.com/glickmac

www.codyglickman.com

Bias in Average Fold Coverage by GC

Average Fold Coverage by GC Content

