Introducing crAssphage
Discovering a virus that is present in half the people in the world!

Rob Edwards
San Diego State University
Metagenomics

Who is in there?

How are they doing it?

What are they doing?
Human Intestinal Viromes

Image: Lisa Brown for National Public Radio
There is no 16S for viruses.

The phage proteomic tree.
This method produces high coverage libraries of over 1 million clones from as little as 1 ng DNA.
200 liters water
5-500 g fresh fecal matter

Concentrate and purify viruses

Extract nucleic acids

DNA/RNA LASL

Sequence

Epifluorescent Microscopy
Most Viral DNA Sequences in Feces are Unknown Phages

- Known: 40%
- Unknown: 60%
- Phages: 94%
- Eukaryotic Viruses: 6%

TBLASTX (E<0.001) ~36,000 sequences

References:
- Breitbart (2002) PNAS
Most Human RNA Viruses are Known

- Known: 92%
- Unknown: 8%
- Pepper Mild Mottle Virus: 65%
- Other Plant Viruses: 9%
- Other: 26%

Pepper Mild Mottle Virus (PMMV)

- ssRNA virus; ≈6 kb genome
- Related to Tobacco Mosaic Virus
- Infects members of *Capsicum* family
- Widely distributed – spread through seeds
- Fruits are small, malformed, mottled
- Rod-shaped virions

![Viral particles in fecal sample](http://www.rothamsted.bbsrc.ac.uk/ppi/links/pplinks/virusems/)

TOBACCO MOSAIC VIRUS
http://www.rothamsted.bbsrc.ac.uk/ppi/links/pplinks/virusems/
PMMV is common in Human Feces

Fecal samples
Extract total RNA
RT-PCR for PMMV

San Diego: 78% people are positive
Singapore: 67% people are positive
México: ?

10-50 fold increase in feces compared to food
$10^6$-$10^9$ PMMV copies per gram dry weight of feces
Which Foods Contain PMMV?

Chili powder
Chili sauces

NOT FOUND IN FRESH PEPPERS

Indian curry
Pork noodle red chili
Chicken rice
Mexican salsa
Hong Kong chili sauce
Hong Kong green chili
Vegetarian chili
Koch’s Postulates

http://www.sweatnspice.com
Most Viral DNA Sequences in Feces are Unknown Phages

Random community genomics

1. Collect mosquitoes
2. Separate or pool mosquitoes
3. Homogenize mosquitoes
4. Filter out eukaryotic cells
5. Density-gradient centrifugation
6. SYBR Gold staining & fluorescence microscopy
7. DNA extraction & amplification
8. Sequencing on GS 20/GS FLX
9. Bioinformatics
FOCUS

Genivaldo Silva
1) Calculate k-mer frequency for the input

2) Find a non-negative vector $x \in \mathbb{R}^n$ to minimize the function

$$f(x) = \frac{1}{2} \|Ax - b\|^2,$$

where $x \geq 0$ and $\sum_{i=1}^{n} x_i = 1$

List of "Known" genomes in the "Unknos"
How long does it take?

- Days to weeks:
  - ~14 days
  - ~3 days

- Minutes to seconds:
  - ~52 mins
  - ~3 mins
  - ~41 secs

Graph showing runtime vs. percentage of oral metagenome (%).
Human microbiome samples

256 GB analyzed in 1 hour and 20 min

Body sites
- saliva
- mid vagina
- supragingival plaque
- attached keratinized gingiva
- buccal mucosa
- subgingival plaque
- vaginal introitus
- throat
- palatine tonsils
- anterior nares
- stool
- right retroauricular crease
- left retroauricular crease
- tongue dorsum
- posterior fornix

Distance from published taxon

[Color scale from 0 to 100]
Bas Dutilh
cross Assembly

<table>
<thead>
<tr>
<th>metagenome 1</th>
<th>metagenome 2</th>
<th>metagenome 3</th>
<th>metagenome 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
cross Assembly

metagenome 1  metagenome 2  metagenome 3  metagenome 4

Assembly
Contigs directly represent the overlap between samples

\[ d_{i,j} = 1 - \frac{1}{2} \cdot \left( \frac{r_{i,j}^2}{r_i^2} + \frac{r_{j,i}^2}{r_j^2} \right) \]

\[ d_{i,j} = \cos^{-1} \left( \sum_{k=1}^{n} \sqrt{p_{ki} \cdot p_{kj}} \right) \]
HMP viruses

Viruses in the faecal microbiota of monozygotic twins and their mothers

Alejandro Reyes, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath, Forest Rohwer & Jeffrey I. Gordon

Reyes et al. Nature 2010
Phages are more variable than microbes

Reyes et al. Nature 2010
*De novo* assembly HMP data

Number of contigs contributing reads to contig

6,988 *de novo* cross-contigs
Coverage of contigs across samples

Samples →

Average depth

1.E-10
1.E-09
1.E-08
1.E-07
1.E-06
1.E-05
1.E-04

→

A
v
e
r
a
g
e
d
e
p
h
→
Complete crAssphage genome
How big is the chimerization problem?

Assembly algorithms include “chimera protection”
- Break contigs at ambiguities

Investigate the effect of chimerization:
- Use different assembly parameters and assess results
  - High stringency → few chimeras
  - Low stringency → many chimeras
What are chimeras?

Chimerization is more frequent between closely related strains

Venus the chimeric cat
https://www.facebook.com/VenusTheAmazingChimeraCat
https://twitter.com/Venustwofacecat

What are intra-phyla chimeras???

0% Photoshopped, 100% Born This Way!
What are chimeras?

Evolutionary conserved entities!

*abundant and conserved enough to assemble*
What is the host?
How to predict the host?

- 2,699 hosts
- 820 phages
- Genetic homology
  - blastn
  - blastx
- CRISPRs
- Exact matches
- GC
- CDS
- $3 \leq k\text{-mer} \leq 8$
- Co-occurrence
- Co-abundance
ROC Curves

blastn

blastx

CRISPRs (species level)

Exact matches

Codon usage

Oligonucleotides (species level)
Number of correct predictions
Predicting crAssphage host

• 1 phage
• ??? host

• Genetic homology
  - blastn
  - blastx

• CRISPRs

• Exact matches
  • GC
  • CDS
  • 3 <= k-mer <=8

• Co-occurrence
• Co-abundance
Protein functions

~40 proteins on forward strand
~40 proteins on reverse strand

Phage structural proteins
DNA manipulation functions

Two *Firmicute* plasmid replication proteins (RepL)

*Bacteroides*-associated carbohydrate binding domains (BACON) on structural protein
Protein homologs

crAssphage is weird among phages!

Co-occurrence of phage and host

- Co-occurrence of phage genome with 404 gut bacteria
  - Reads mapped from 152 fecal total community metagenomes
  - Normalize; Spearman rank correlations; cluster
Co-occurrence of phage and host

crAssphage clusters with Bacteroidetes

Just like two known Bacteroides phages B40-8 and B124-14
Plaque assays

- Requires correct host strain
- Requires phage makes visible plaques
- Often requires correct concentrations of Mg$^{++}$, Ca$^{++}$, etc

No PCR hits in at least 100 plaques isolated from 10 pooled viral preparations on *Bacteroides fragilis* and *B. thetaiotaomicron* lawns.
Where is crAssphage found?
crAssphage found in intestines

Looked at 2,906 metagenomes
Only found in 940 metagenomes
(all human associated)
crAssphage is abundant!

Abundance-ubiquity plot

- crAssphage
- PhAnToMe phages
crAssphage by the numbers

• Present in 32.3% of sequenced environmental samples (940 / 2,906)
  – Includes virus metagenomes and total community metagenomes

• >6x more abundant than all (1,192) other known phages combined
  – Corrected for genome size

• Present in 73.4% of sequenced human fecal samples (342 / 466)
  – 99.9% of all crAssphage reads were found in feces (significant)

• 1.68% of the reads in all human fecal metagenomes

• Estimate: ~6 crAssphage genomes per Bacteroides genome in your gut

• >90% of the reads in some of virus metagenomes from the US twin study
• 24% of the reads in an unrelated virus metagenome from Korea
• 22% of the reads in total community metagenomes from USA (HMP data)
• Found on every continent (where we have data)
Virome reads mapping to viral database

Viral database vs crAssphage

Reyes et al. Nature 2010
Potential caveats

- Phage or contamination?
  - Highly abundant in viral metagenomes size- and density-filtered for VLPs
  - ORFs show similarity to bacteriophage and bacterial proteins (no conserved bacterial or archaeal metabolic genes)
  - Phage-like modularity among functions
  - Coding structure of the ORFs is typical of a phage genome
  - Putative prokaryotic promoter patterns
  - Genome detected in many metagenomes around the world

- Amplification skews?
Summary

• crAssphage is everywhere
• everyone has it (rounding up)
• we don't know what it does
Funding

PhAnToMe
TUES
Viral Dark Matter
Phenotype-Genotype
Big computers

Brazil-US Marine Sciences Consortium

Coral Reef Image Analysis

Fund for the Improvement of Postsecondary Education (FIPSE) U.S. Department of Education.