Viromics: what, how, why?

Evelien Adriaenssens
Andrea Telatin

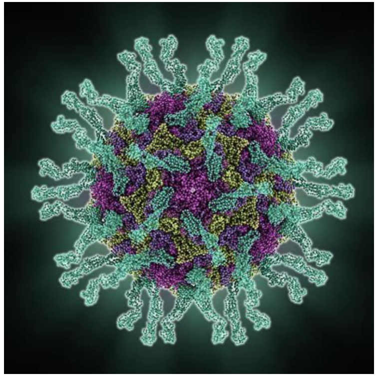
Quadram Institute
Science Health Food Innovation

What is a virus?

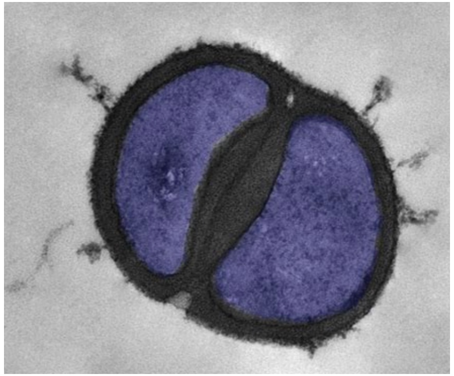
Quadram Institute
Science Health Food Innovation

A virus or not a virus?

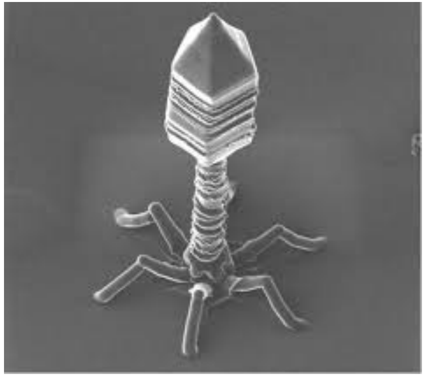
A



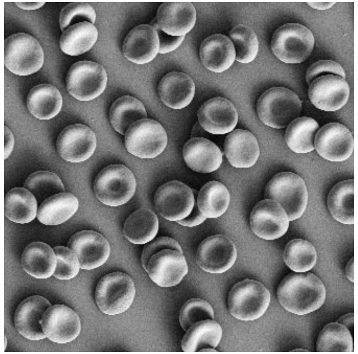
B



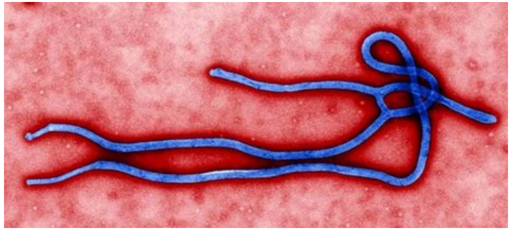
C



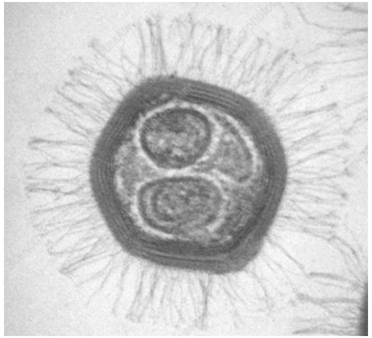
D



E



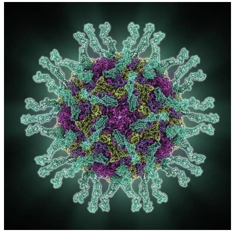
F



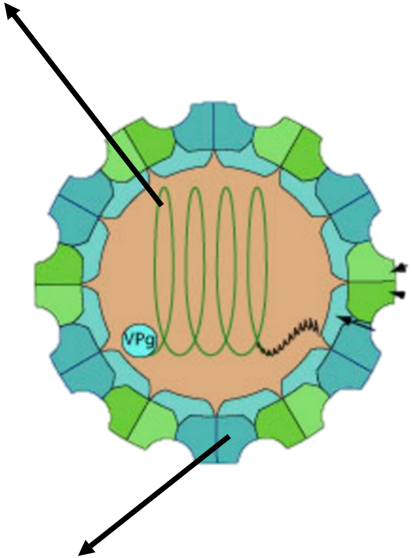
G

https://tinyurl.com/ebame-virus

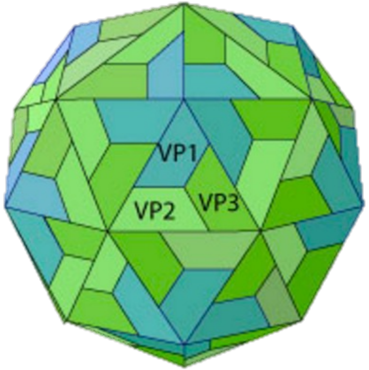
Example A: poliovirus



genetic information as single segment of RNA



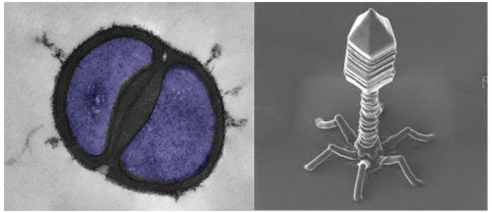
capsid or coat made of different proteins

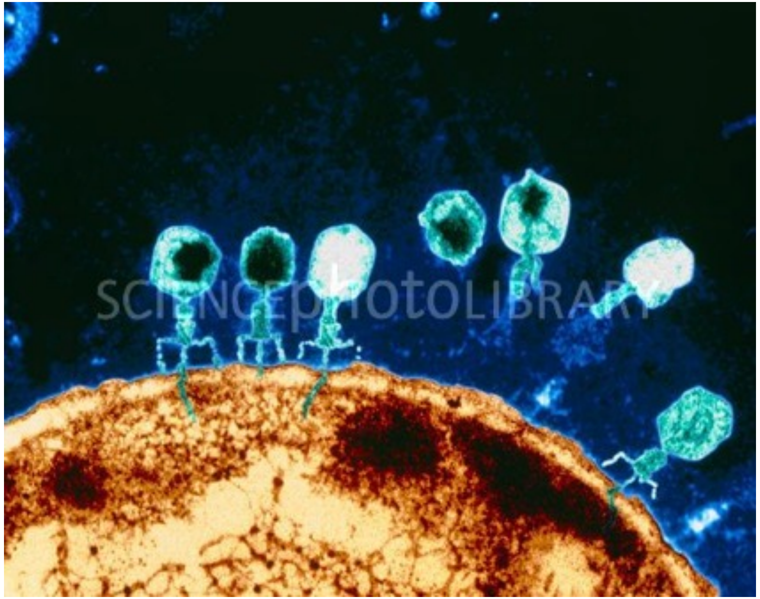


(+)ssRNA molecule
enteroviruses 7-8.5 kb
polyprotein

viralzone.expasy.org

Examples B & C: Staphylococcus cell and bacteriophage

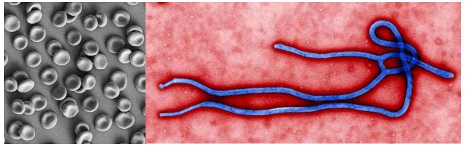


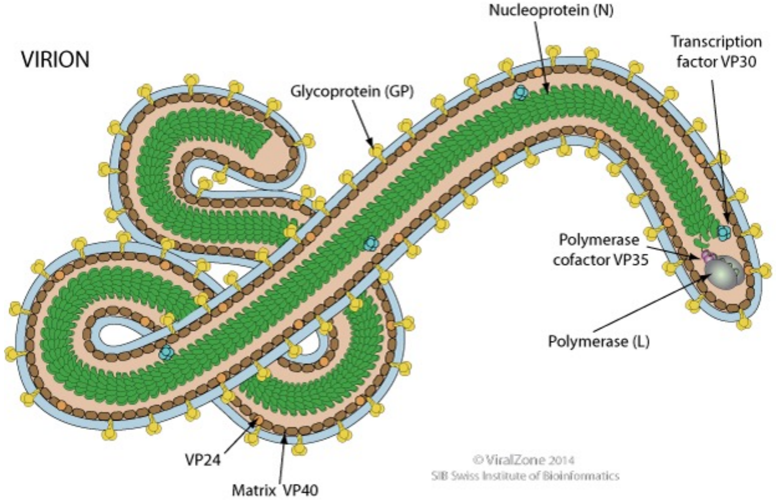


virus = bacteriophage
most commonly detected virus in metagenomics

Tailed bacteriophages (*Caudoviricetes*)
protein coat + genetic material (dsDNA)
linear dsDNA molecules
can be circularly permuted, have cohesive overhangs, or terminal repeats
~15 - > 500 kb

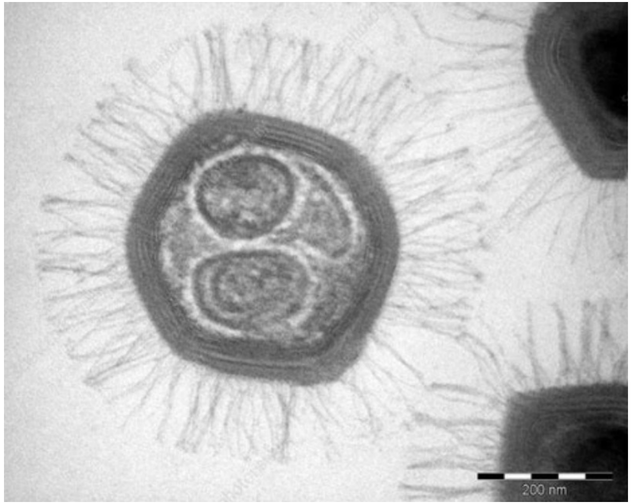
Examples D & E: red blood cells and ebolavirus

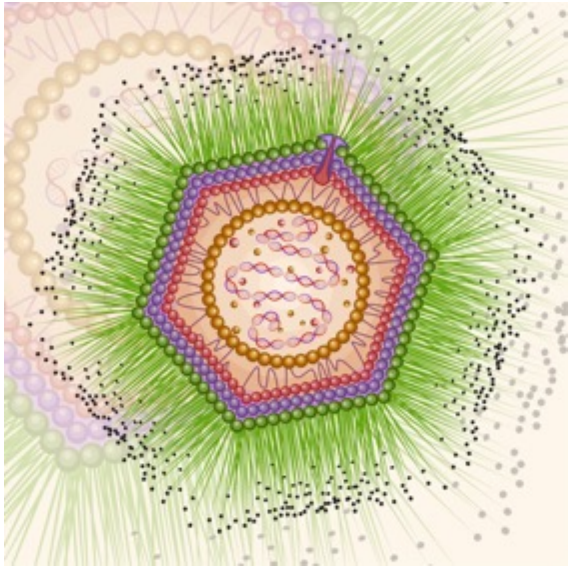




very large virion (970 nm long)
(-)ssRNA genome
18-19 kb

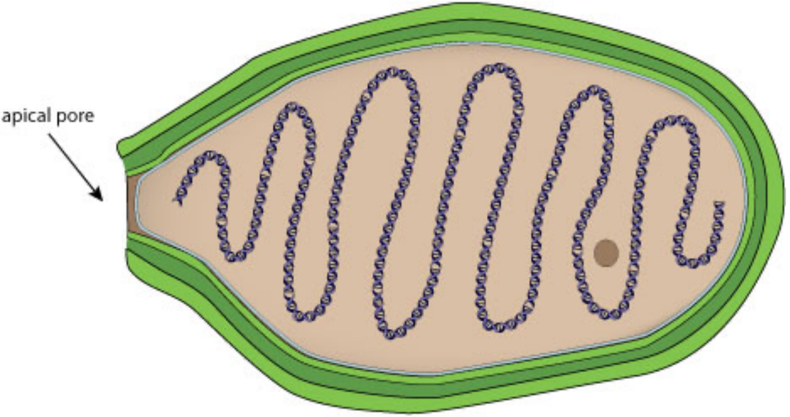
Example F: Mimivirus





Giant virus structure (visible with normal light microscope)
dsDNA genome, linear molecule, *Mimiviridae* genomes 350 kb – 1.2 Mb

Example G: Pandoravirus



© ViralZone 2014
SIB Swiss Institute of Bioinformatics

particle 1 µm long

Pandoravirus salinus has the largest known virus genome: 2.4 Mb

linear dsDNA

How does this diversity impact viromics or viral metagenomics?

Problems!

Everything you do creates a bias

Different particle sizes: 20 nm - > 1µm

Different capsid structures

Different genome types: ssDNA, dsDNA, (+/-)ssRNA, dsRNA

Different genome sizes: ~1,300 - > 2,000,000 nt

Segmented genomes

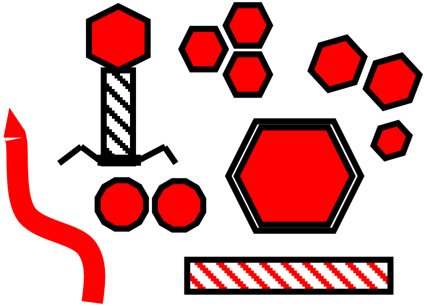
Linear or circular

No conserved marker gene across the virosphere

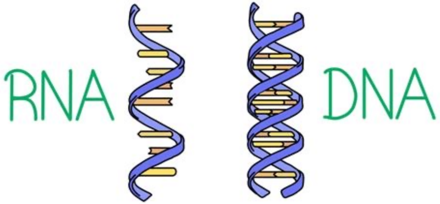
No shared evolutionary origin

BUT, majority of viruses <200 nm and genomes <200 kb

Viromics

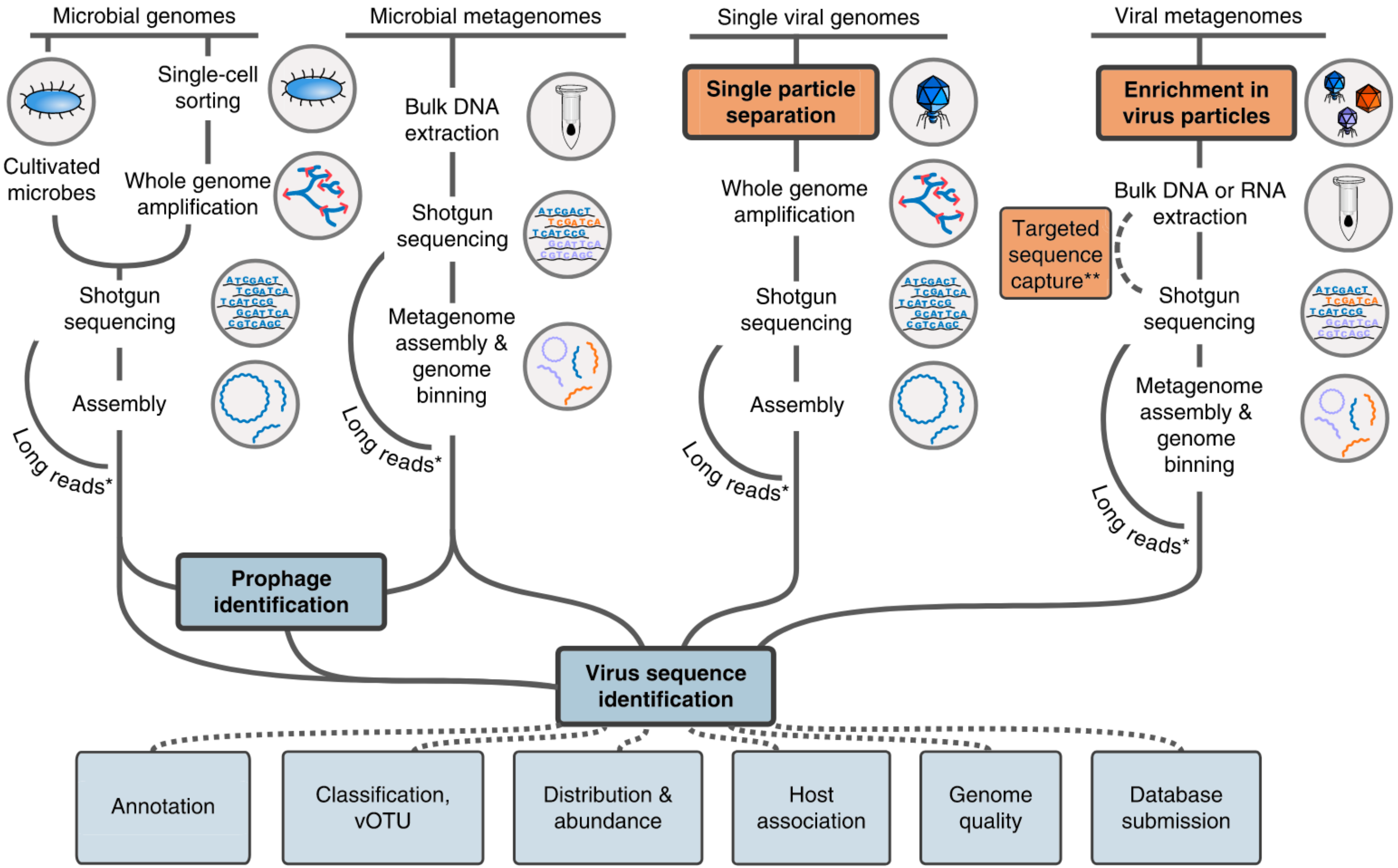


separate virus community
from cellular material



nucleic acid extraction

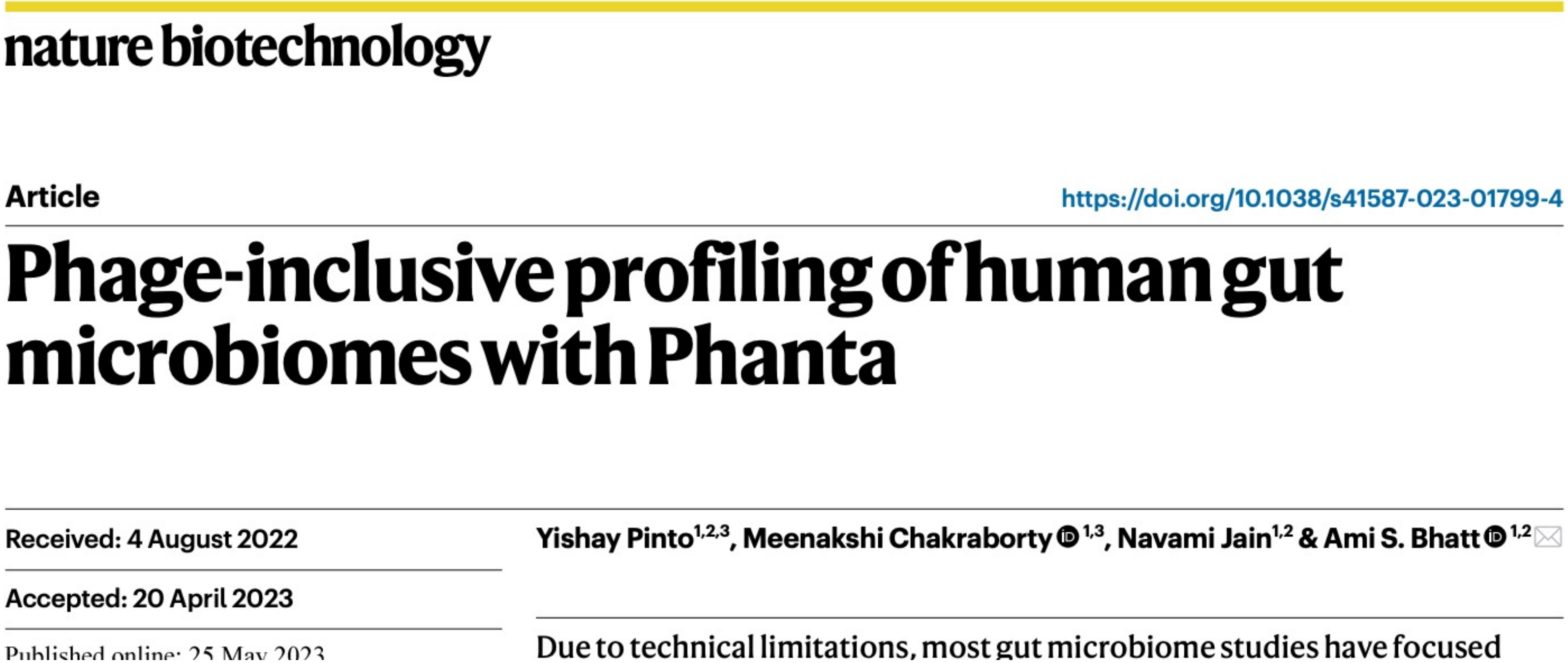
sequencing



Roux et al, Nature Biotechnology, 2019, Minimum Information about an Uncultivated Virus Genome

Quick & “dirty” analyses: read profiling

Quadram Institute
Science Health Food Innovation



nature biotechnology

Article https://doi.org/10.1038/s41587-023-01799-4

Phage-inclusive profiling of human gut microbiomes with Phanta

Received: 4 August 2022

Accepted: 20 April 2023

Yishay Pinto[1,2,3], Meenakshi Chakraborty[1,3], Navami Jain[1,2] & Ami S. Bhatt[1,2]

Due to technical limitations, most gut microbiome studies have focused

nature biotechnology

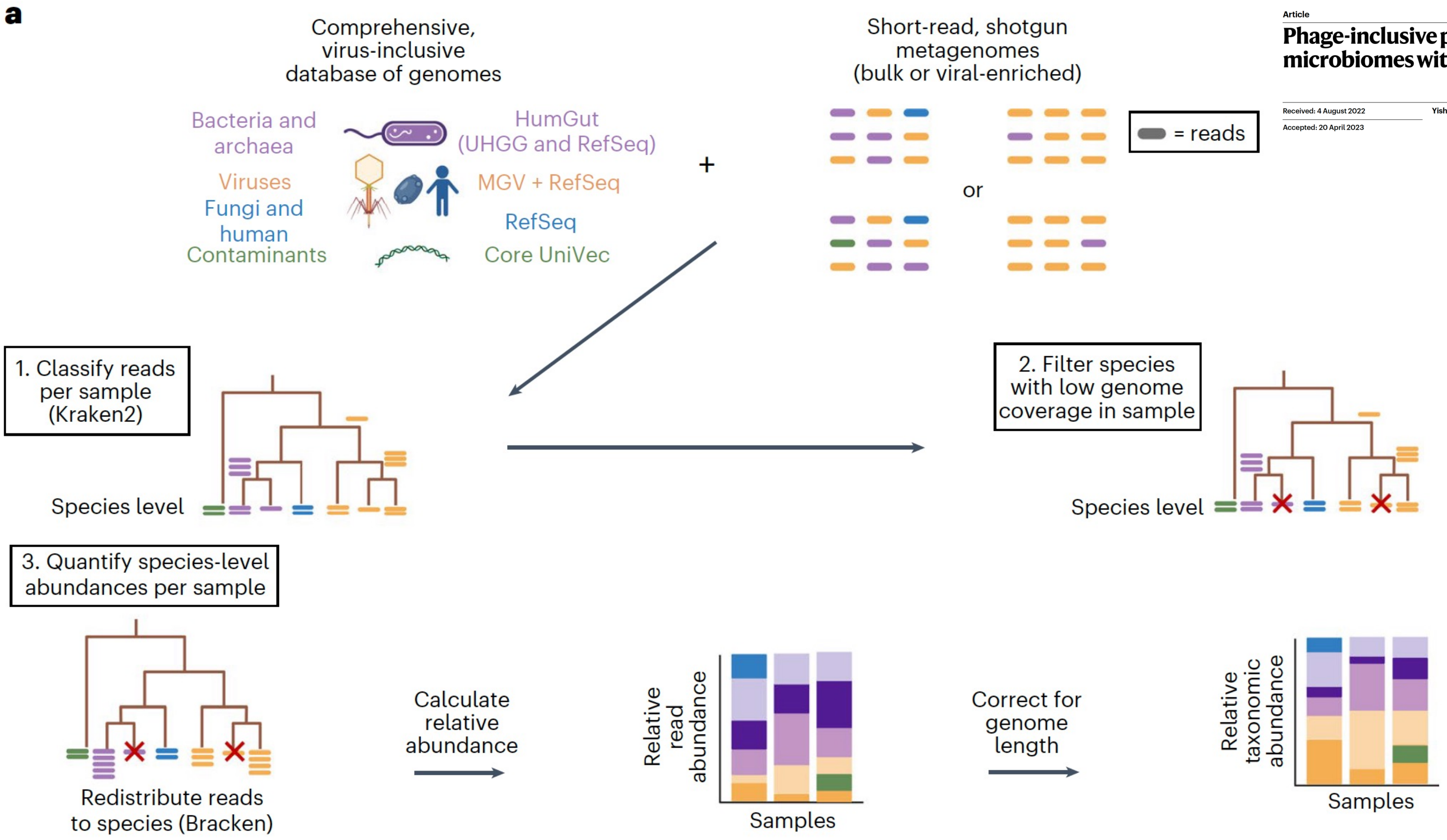
Article

https://doi.org/10.1038/s41587-023-01799-4

Phage-inclusive profiling of human gut microbiomes with Phanta

Received: 4 August 2022
Accepted: 20 April 2023

Yishay Pinto[1,2,3], Meenakshi Chakraborty[1,3], Navami Jain[1,2] & Ami S. Bhatt[1,2] ✉



Comprehensive, virus-inclusive database of genomes

Bacteria and archaea — HumGut (UHGG and RefSeq)

Viruses — MGV + RefSeq

Fungi and human — RefSeq

Contaminants — Core UniVec

+

Short-read, shotgun metagenomes (bulk or viral-enriched)

or

= reads

1. Classify reads per sample (Kraken2)

Species level

2. Filter species with low genome coverage in sample

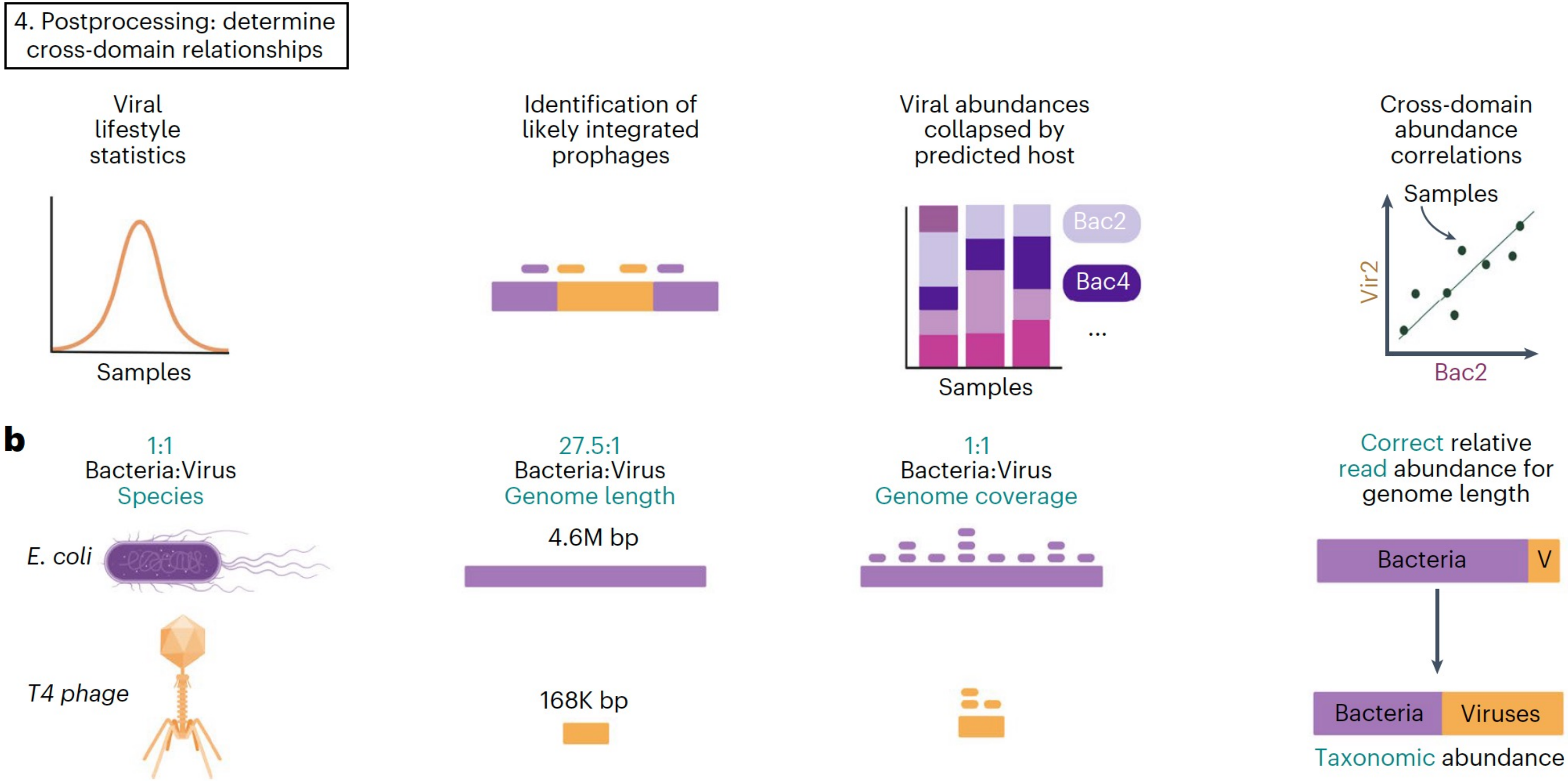
Species level

3. Quantify species-level abundances per sample

Redistribute reads to species (Bracken)

Calculate relative abundance

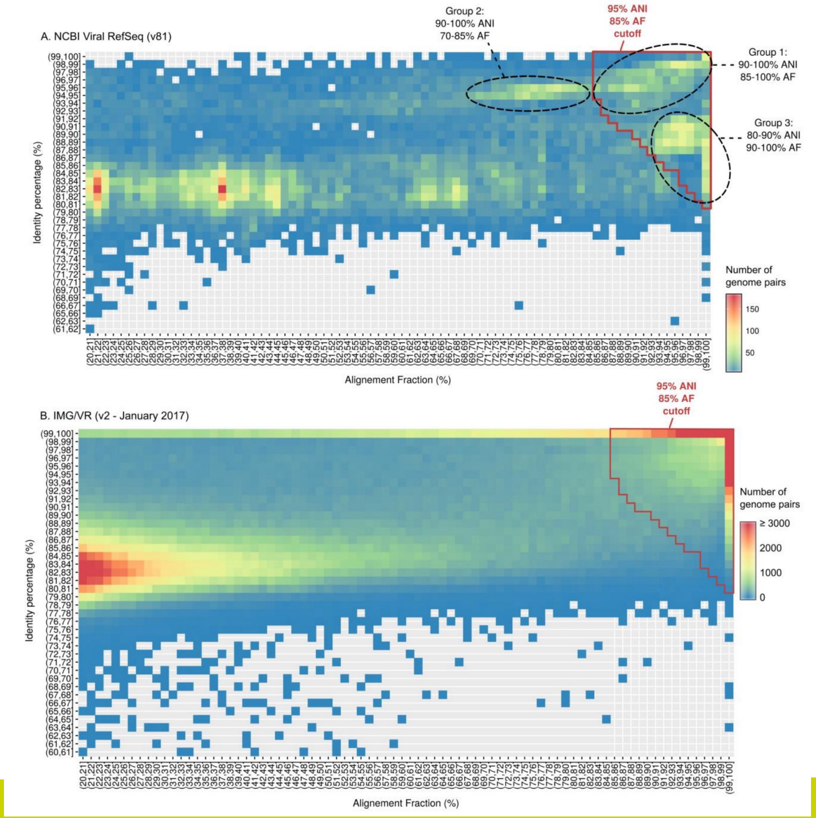
Correct for genome length

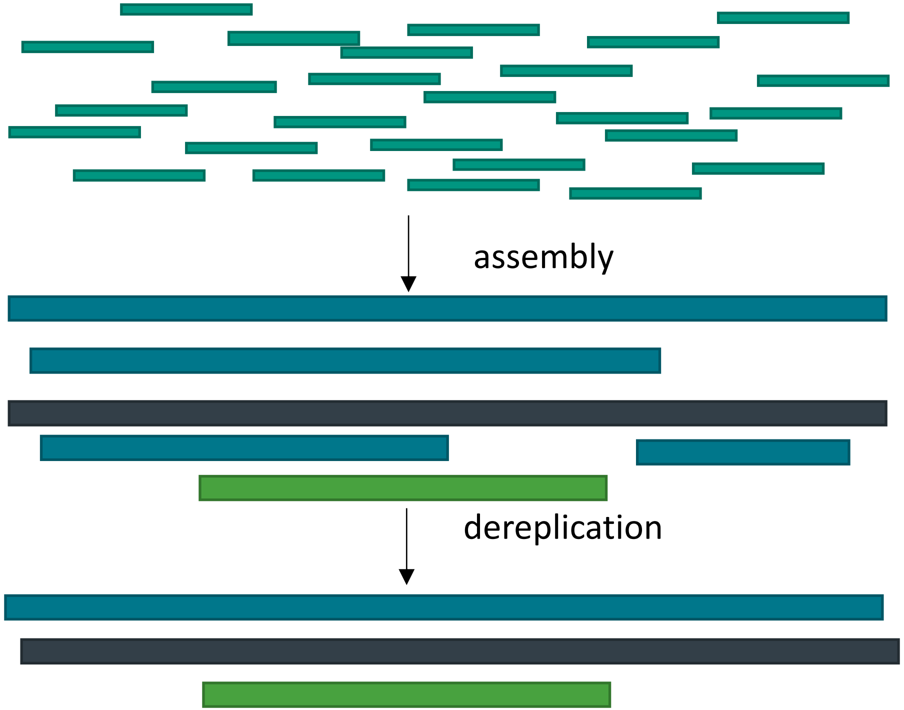

4. Postprocessing: determine cross-domain relationships
Viral lifestyle statistics
Samples
Identification of likely integrated prophages
Viral abundances collapsed by predicted host
Bac2
Bac4
...
Samples
Cross-domain abundance correlations
Samples
Vir2
Bac2
b
1:1
Bacteria:Virus
Species
E. coli
T4 phage
27.5:1
Bacteria:Virus
Genome length
4.6M bp
168K bp
1:1
Bacteria:Virus
Genome coverage
Correct relative read abundance for genome length
Bacteria
V
Bacteria
Viruses
Taxonomic abundance

Gold standard: vOTU approach

Quadram Institute
Science Health Food Innovation

What is a vOTU?





Viromes:

>95% ANI over >85% of alignment: vOTU

~ tailed phage species

Roux et al, Nature Biotechnology, 2019, Minimum Information about an Uncultivated Virus Genome

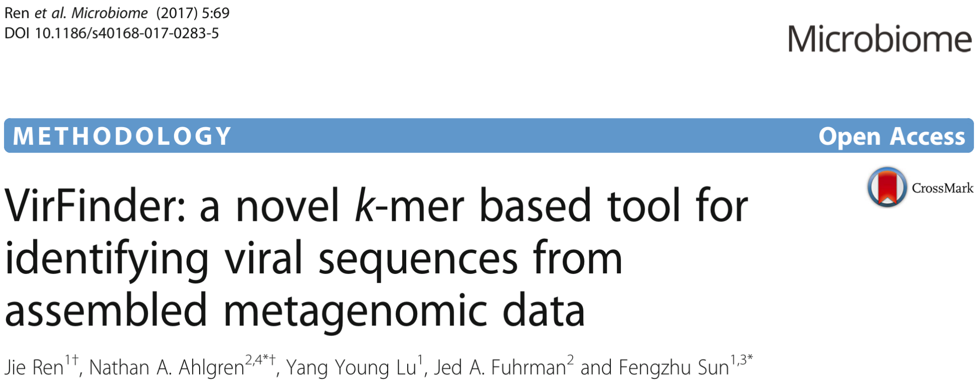
What if not all contigs are viral?

➔ virus mining

VirSorter: mining viral signal from microbial genomic data

Simon Roux[1,*], Francois Enault[2,3], Bonnie L. Hurwitz[4], Matthew B. Sullivan[1,*]

Published May 28, 2015



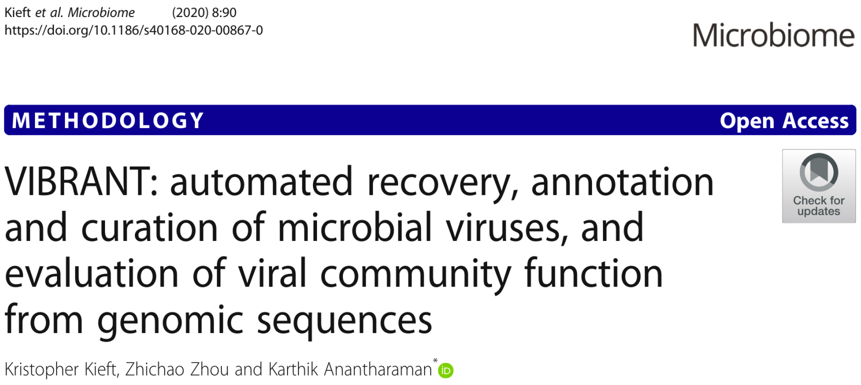
Ren *et al. Microbiome* (2017) 5:69
DOI 10.1186/s40168-017-0283-5

Microbiome

METHODOLOGY Open Access

VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data

Jie Ren[1†], Nathan A. Ahlgren[2,4*†], Yang Young Lu[1], Jed A. Fuhrman[2] and Fengzhu Sun[1,3*]



Kieft *et al. Microbiome* (2020) 8:90
https://doi.org/10.1186/s40168-020-00867-0

Microbiome

METHODOLOGY Open Access

VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences

Kristopher Kieft, Zhichao Zhou and Karthik Anantharaman[*]

Quantitative Biology 2020, 8(1): 64–77
https://doi.org/10.1007/s40484-019-0187-4

RESEARCH ARTICLE

Identifying viruses from metagenomic data using deep learning

Jie Ren[1,†,*], Kai Song[2,†], Chao Deng[1], Nathan A. Ahlgren[3], Jed A. Fuhrman[4], Yi Li[5], Xiaohui Xie[5], Ryan Poplin[6], Fengzhu Sun[1,*]

Guo *et al. Microbiome* (2021) 9:37
https://doi.org/10.1186/s40168-020-00990-y

Microbiome

SOFTWARE ARTICLE Open Access

VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses

Jiarong Guo[1], Ben Bolduc[1], Ahmed A. Zayed[1], Arvind Varsani[2,3], Guillermo Dominguez-Huerta[1], Tom O. Delmont[4], Akbar Adjie Pratama[1], M. Consuelo Gazitúa[5], Dean Vik[1], Matthew B. Sullivan[1,6,7*] and Simon Roux[8*]

nature biotechnology

Article https://doi.org/10.1038/s41587-023-01953-y

Identification of mobile genetic elements with geNomad

Received: 6 March 2023

Accepted: 17 August 2023

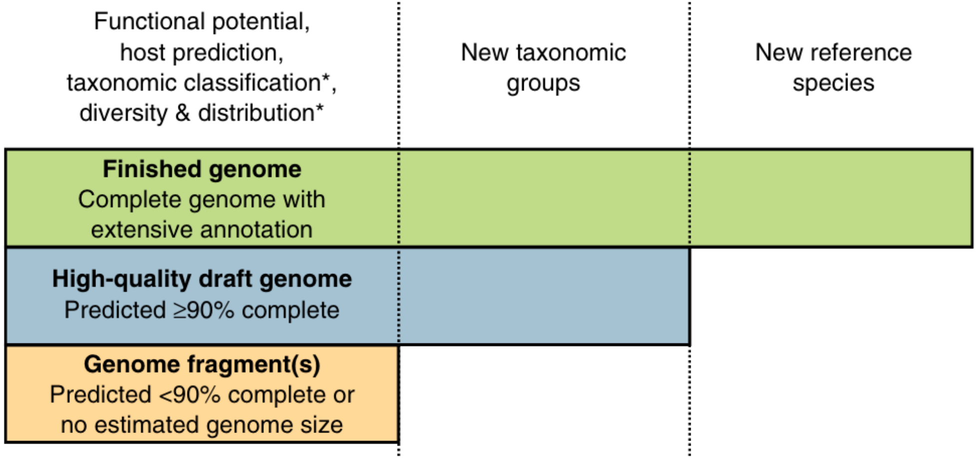
Antonio Pedro Camargo[1] ✉, Simon Roux[1], Frederik Schulz[1], Michal Babinski[2], Yan Xu[2], Bin Hu[2], Patrick S. G. Chain[2], Stephen Nayfach[1] & Nikos C. Kyrpides[1] ✉

input data: assemblies

→ virus mining magic

outputs:
list of contigs
list of probable contigs
fasta files with predictions

How do we know the virus genomes are complete?



Functional potential, host prediction, taxonomic classification*, diversity & distribution*

New taxonomic groups

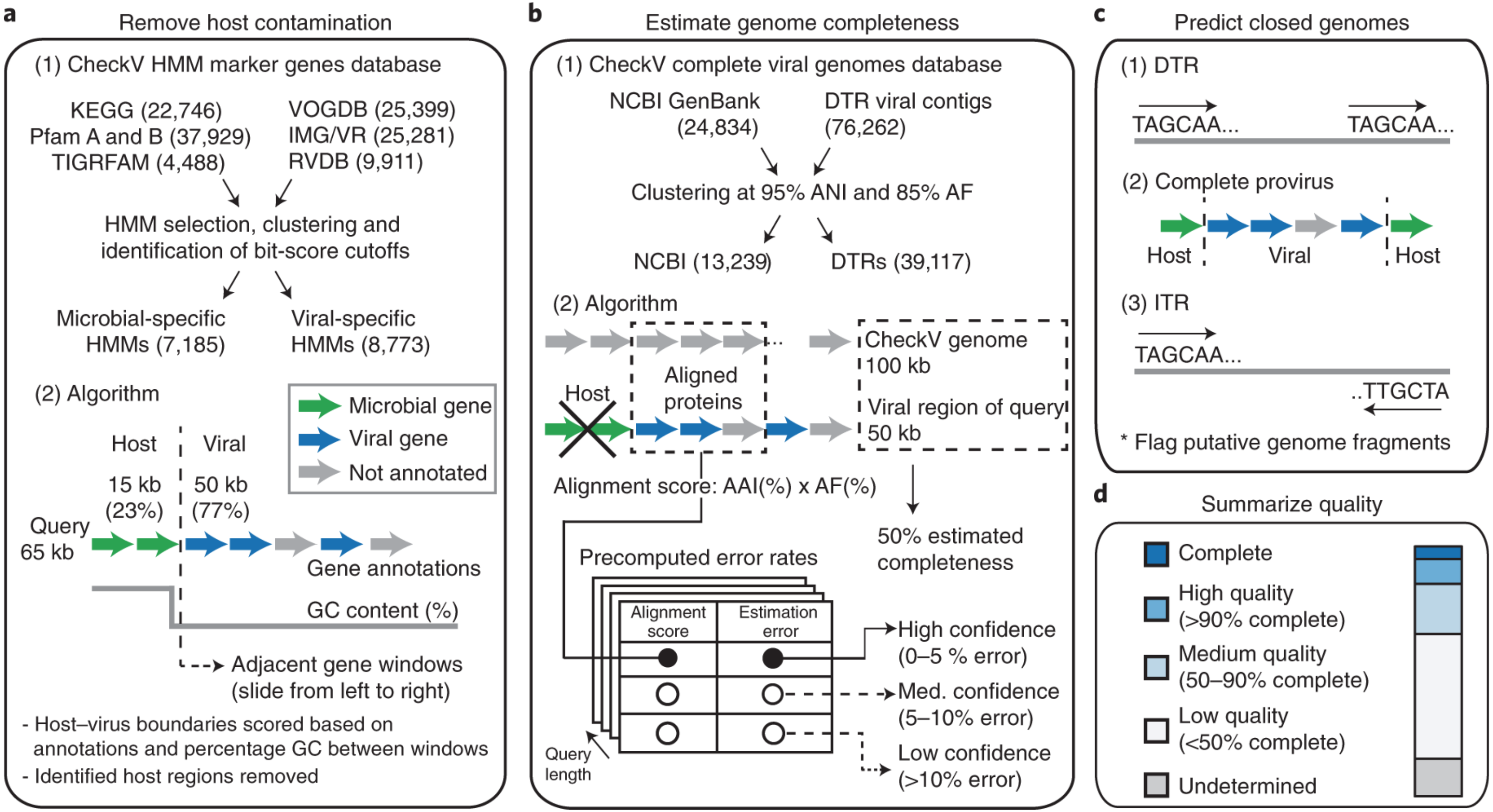
New reference species

Finished genome
Complete genome with extensive annotation

High-quality draft genome
Predicted ≥90% complete

Genome fragment(s)
Predicted <90% complete or no estimated genome size

Roux et al, Nature Biotechnology, 2019, Minimum Information about an Uncultivated Virus Genome



CheckV

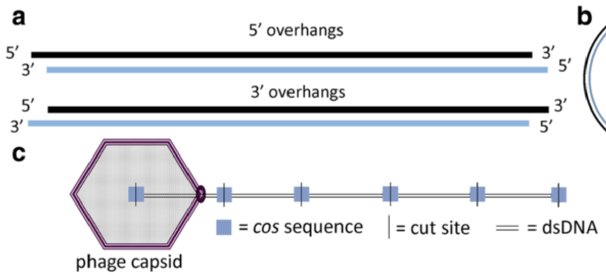
https://bitbucket.org/berkeleylab/checkv/src/master/

Nayfach et al, 2020, Nature Biotechn

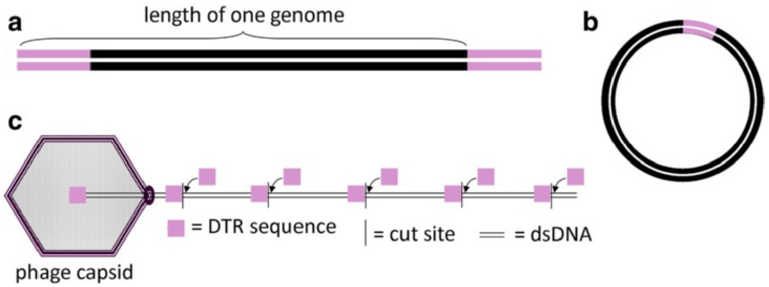
Quadram Institute
Science Health Food Innovation

Phage genome examples: defined ends

Cohesive ends



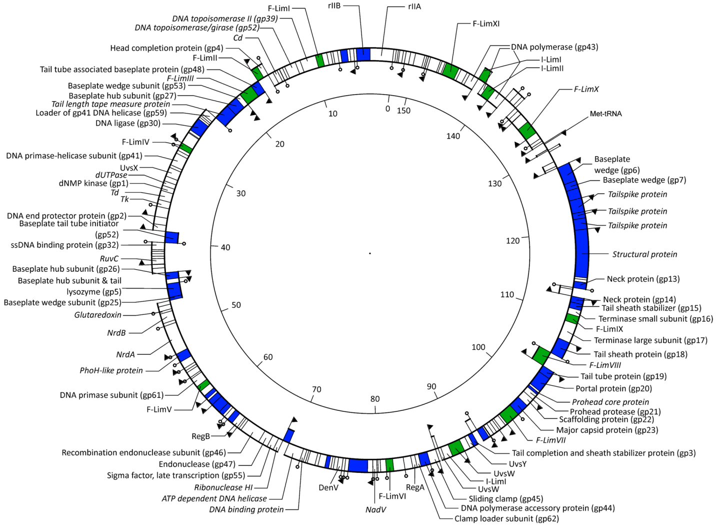
Terminal repeats

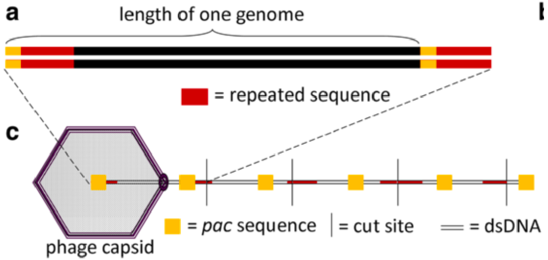


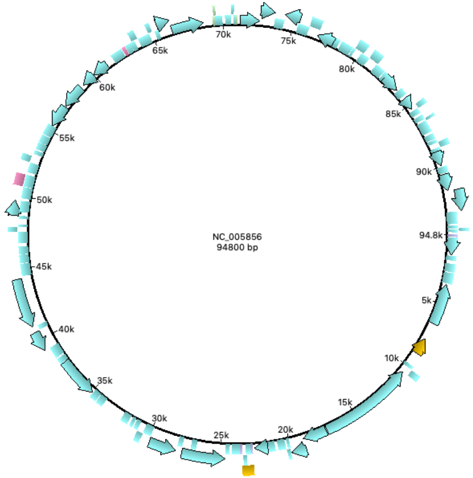
Phage genome examples: circularly permuted genomes

headful packaging: random ends

pac sites: start fixed, end random





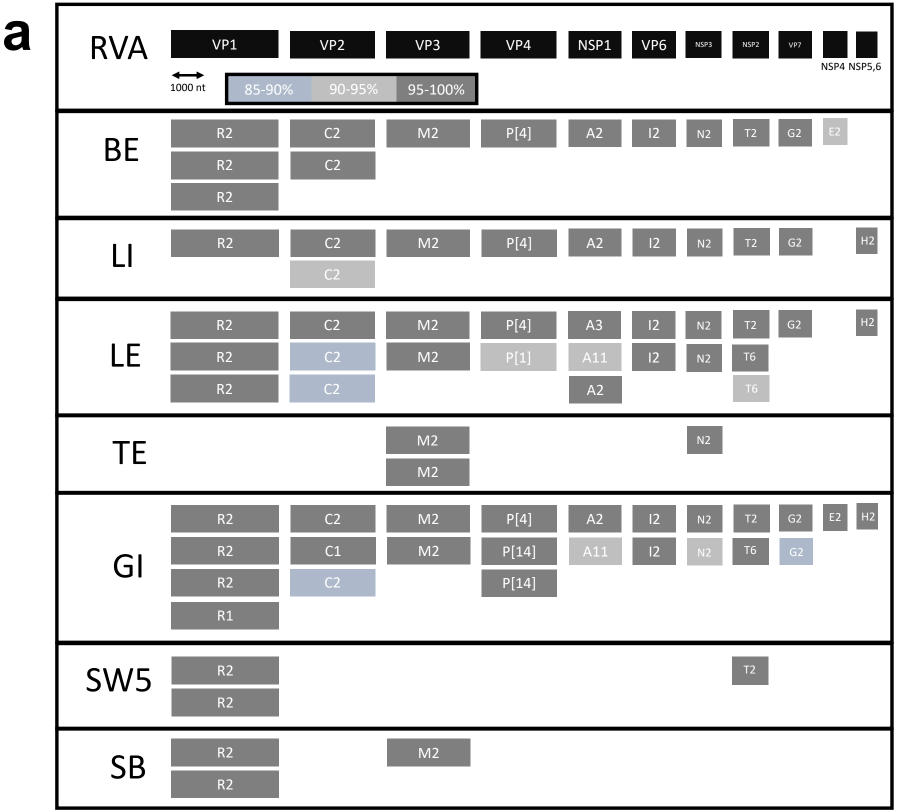


Merrill et al, 2016, BMC Genomics

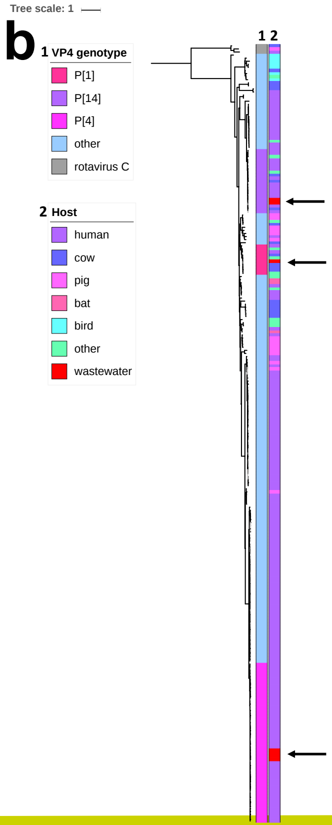
Dickeya phage LIMEstone1, Adriaenssens et al 2012, PLoS ONE

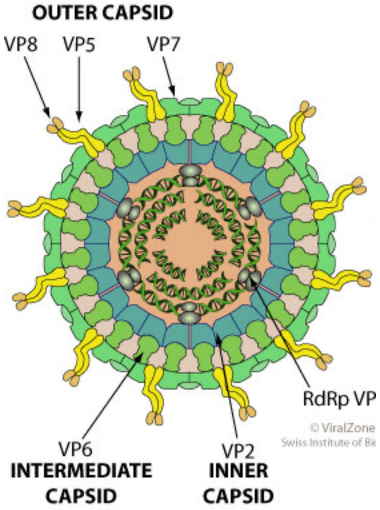
Escherichia phage P1, Lobocka et al 2004, J. Bact.

Segmented genomes: what the...?



RVA	VP1	VP2	VP3	VP4	NSP1	VP6	NSP3	NSP2	VP7	NSP4	NSP5,6
BE	R2 R2 R2	C2 C2	M2	P[4]	A2	I2	N2	T2	G2	E2	
LI	R2	C2 C2	M2	P[4]	A2	I2	N2	T2	G2		H2
LE	R2 R2 R2	C2 C2 C2	M2 M2	P[4] P[1]	A3 A11 A2	I2 I2	N2 N2	T2 T6 T6	G2		H2
TE			M2 M2				N2				
GI	R2 R2 R2 R1	C2 C1 C2	M2 M2	P[4] P[14] P[14]	A2 A11	I2 I2	N2 N2	T2 T6	G2 G2	E2	H2
SW5	R2 R2							T2			
SB	R2 R2		M2								

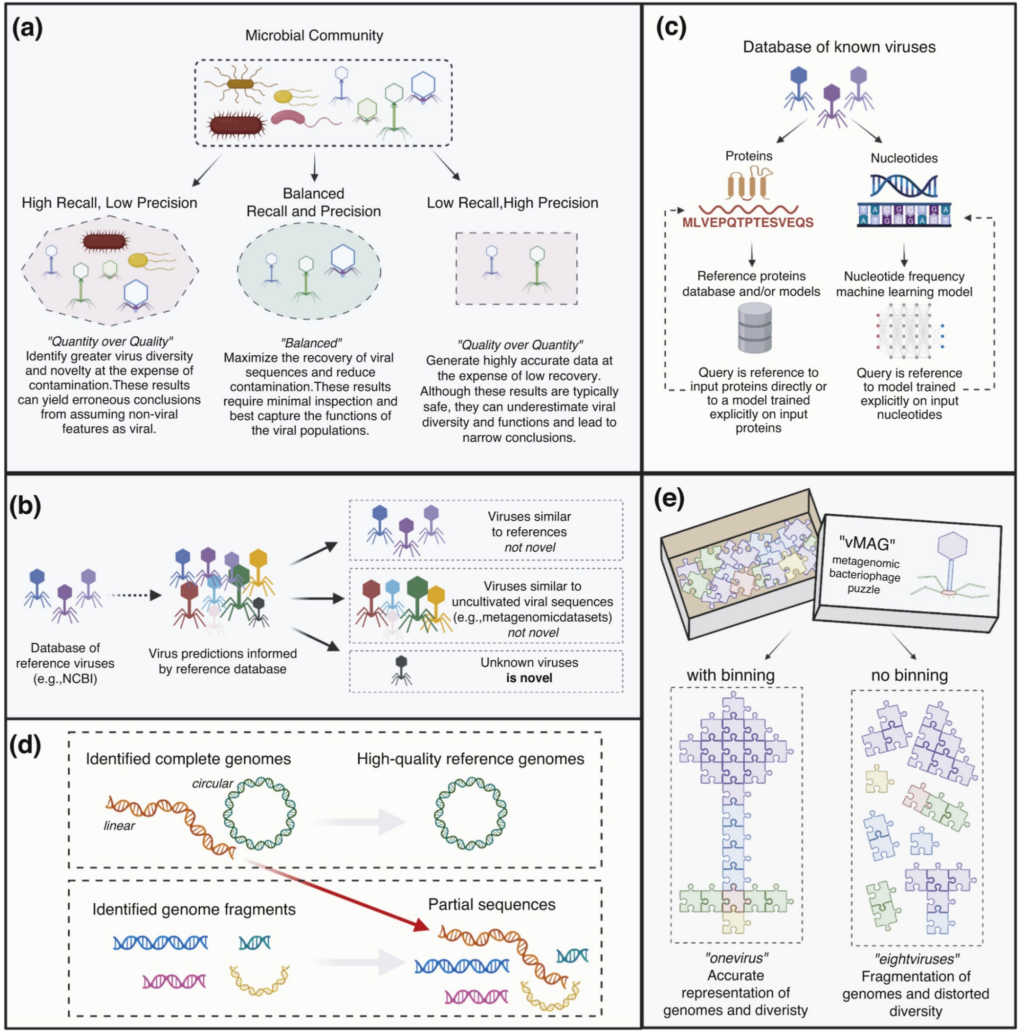




Adriaenssens et al, 2021, Water Research

No tool is perfect!

Bias is everywhere



Current Opinion in Virology

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Virology

Virus genomics: what is being overlooked?

Kristopher Kieft[1,2] and Karthik Anantharaman[1]

Kieft & Anantharaman, 2022

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

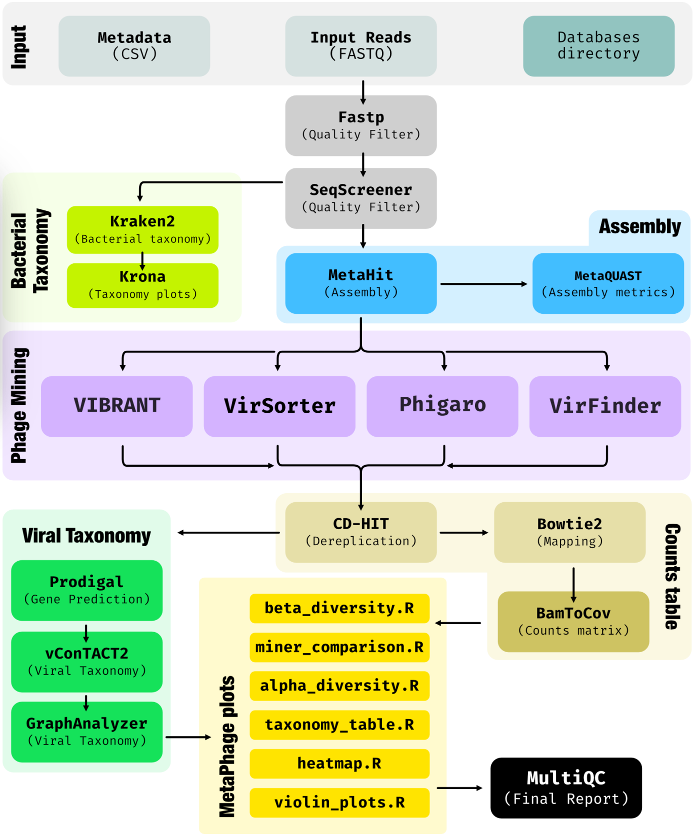
Open Peer Review | Bacteriophages | Methods and Protocols | 7 September 2022

MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data

Authors: Mattia Pandolfo, Andrea Telatin, Gioele Lazzari, Evelien M. Adriaenssens, Nicola Vitulo | AUTHORS INFO & AFFILIATIONS

A reads-to-report pipeline for the analysis of the viral component of metagenomes

https://mattiapandolfovr.github.io/MetaPhage



Input

Metadata (CSV)

Input Reads (FASTQ)

Databases directory

Fastp (Quality Filter)

SeqScreener (Quality Filter)

Bacterial Taxonomy

Kraken2 (Bacterial taxonomy)

Krona (Taxonomy plots)

Assembly

MetaHit (Assembly)

MetaQUAST (Assembly metrics)

Phage Mining

VIBRANT

VirSorter

Phigaro

VirFinder

Viral Taxonomy

Prodigal (Gene Prediction)

vConTACT2 (Viral Taxonomy)

GraphAnalyzer (Viral Taxonomy)

CD-HIT (Dereplication)

Bowtie2 (Mapping)

BamToCov (Counts matrix)

Counts table

MetaPhage plots

beta_diversity.R

miner_comparison.R

alpha_diversity.R

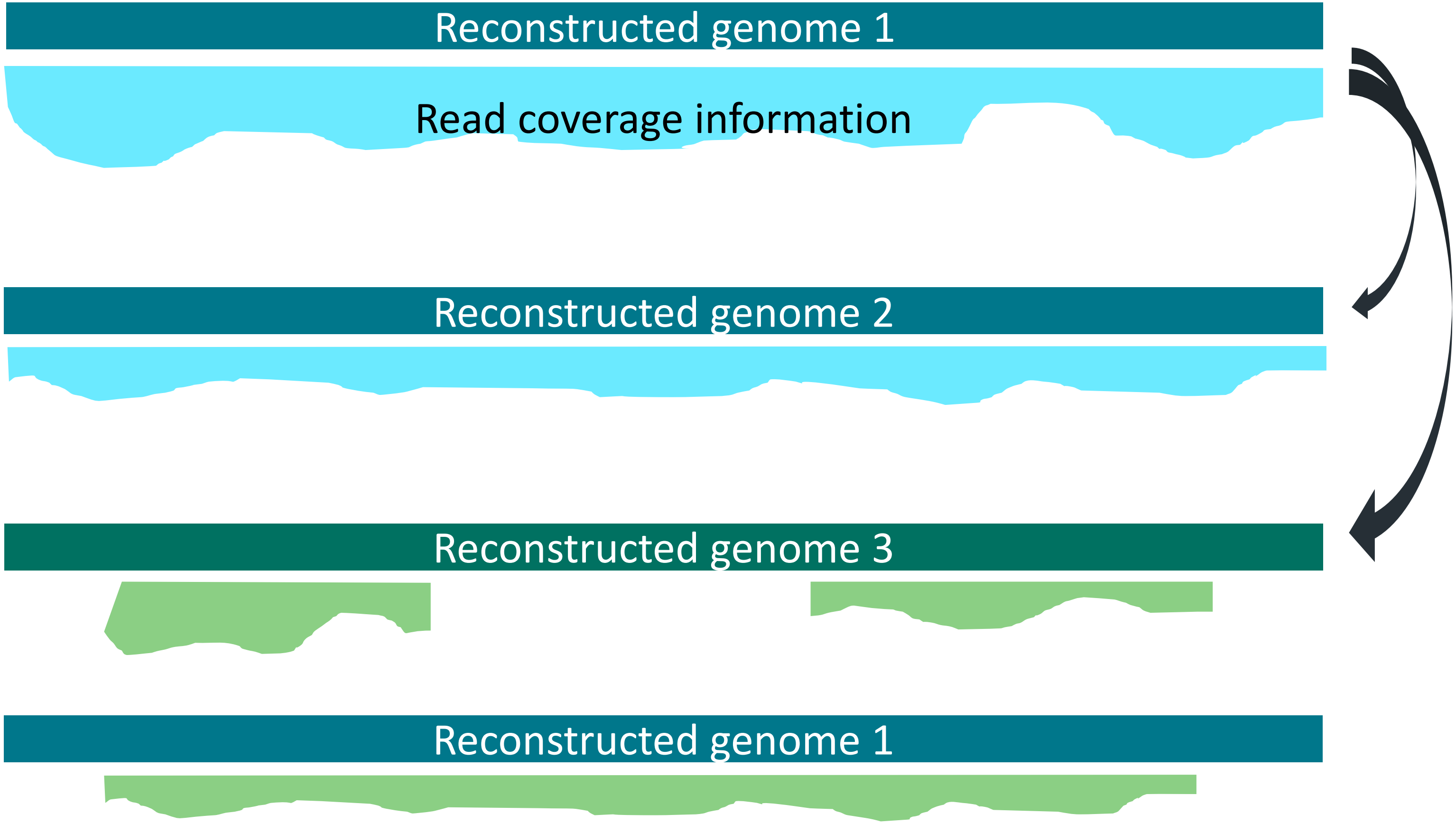
taxonomy_table.R

heatmap.R

violin_plots.R

MultiQC (Final Report)

Quadram Institute
Science · Health · Food · Innovation



Reconstructed genome 1

Read coverage information

Reconstructed genome 2

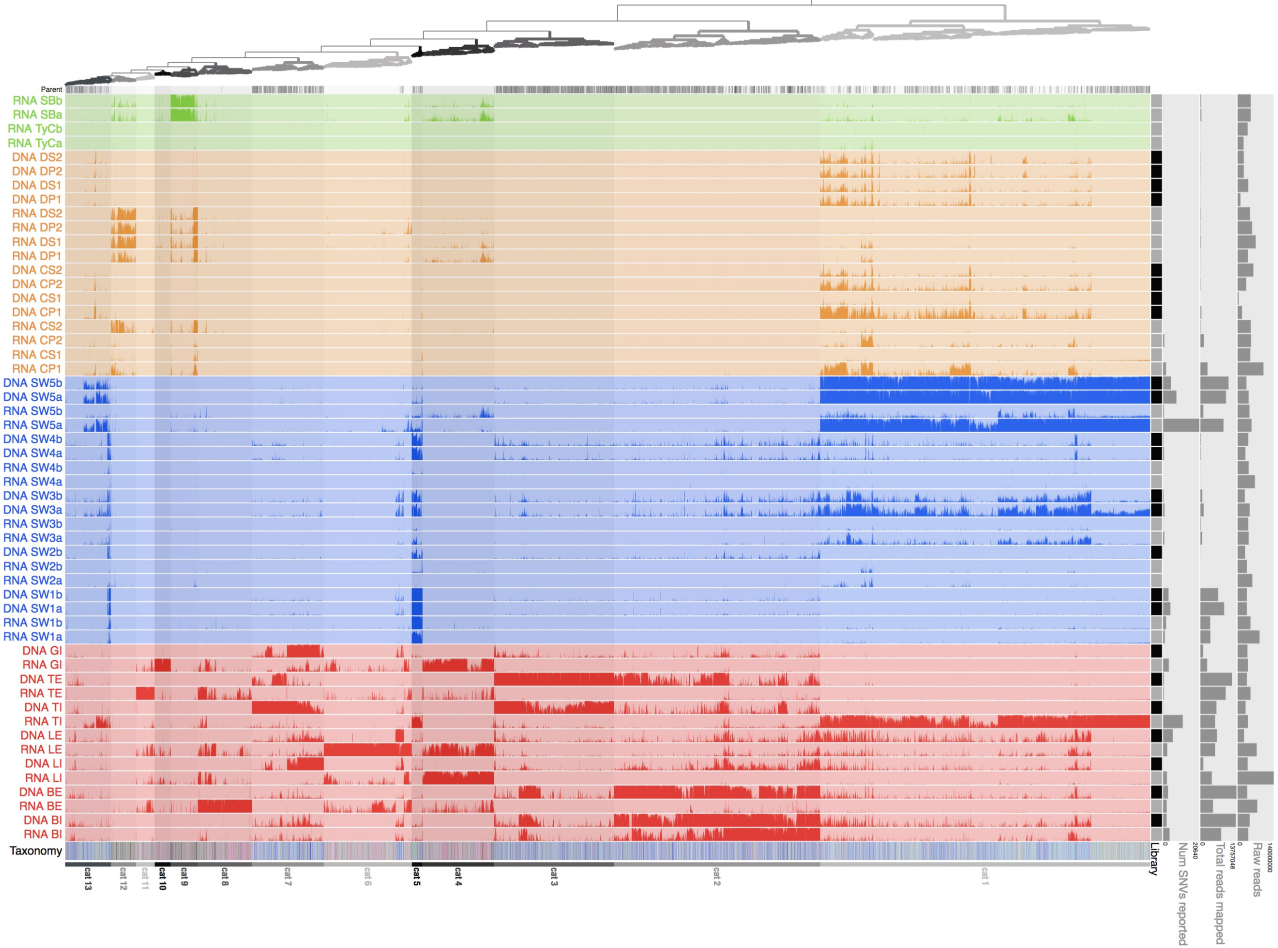
Reconstructed genome 3

Reconstructed genome 1

Compare abundance

Decision on detection

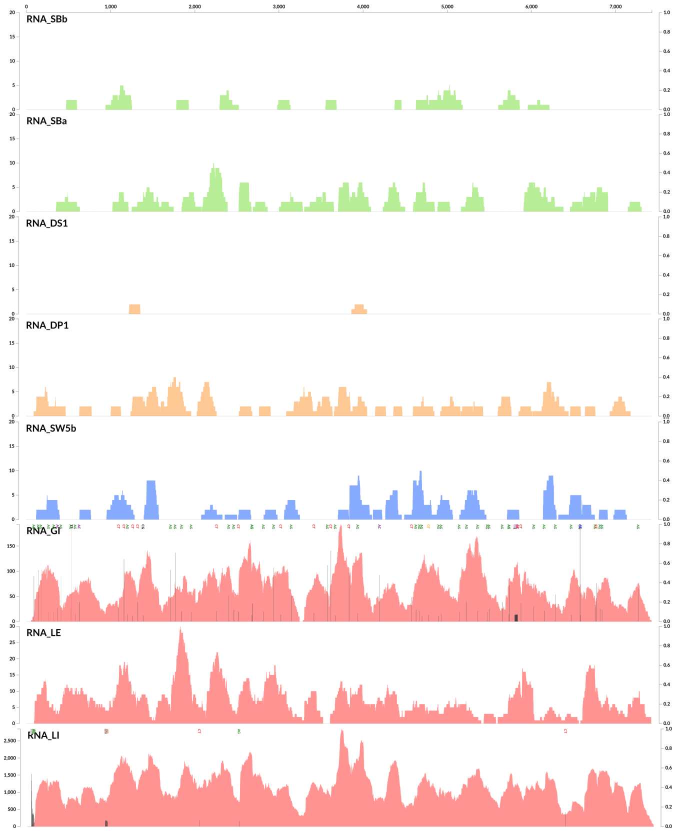
Can we use these outputs in anvi’o?



Adriaenssens et al, 2021, Water Research: analysis with Anvi’o v5 and v6

LI_NODE_9_length_7431_cov_255.051_split_00001 detailed

Sapovirus



Outputs so far:

- viral profile from reads
- set of viral contigs/genomes
- estimates of completeness
- relative abundances

Quadram Institute
Science Health Food Innovation