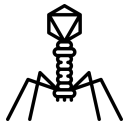
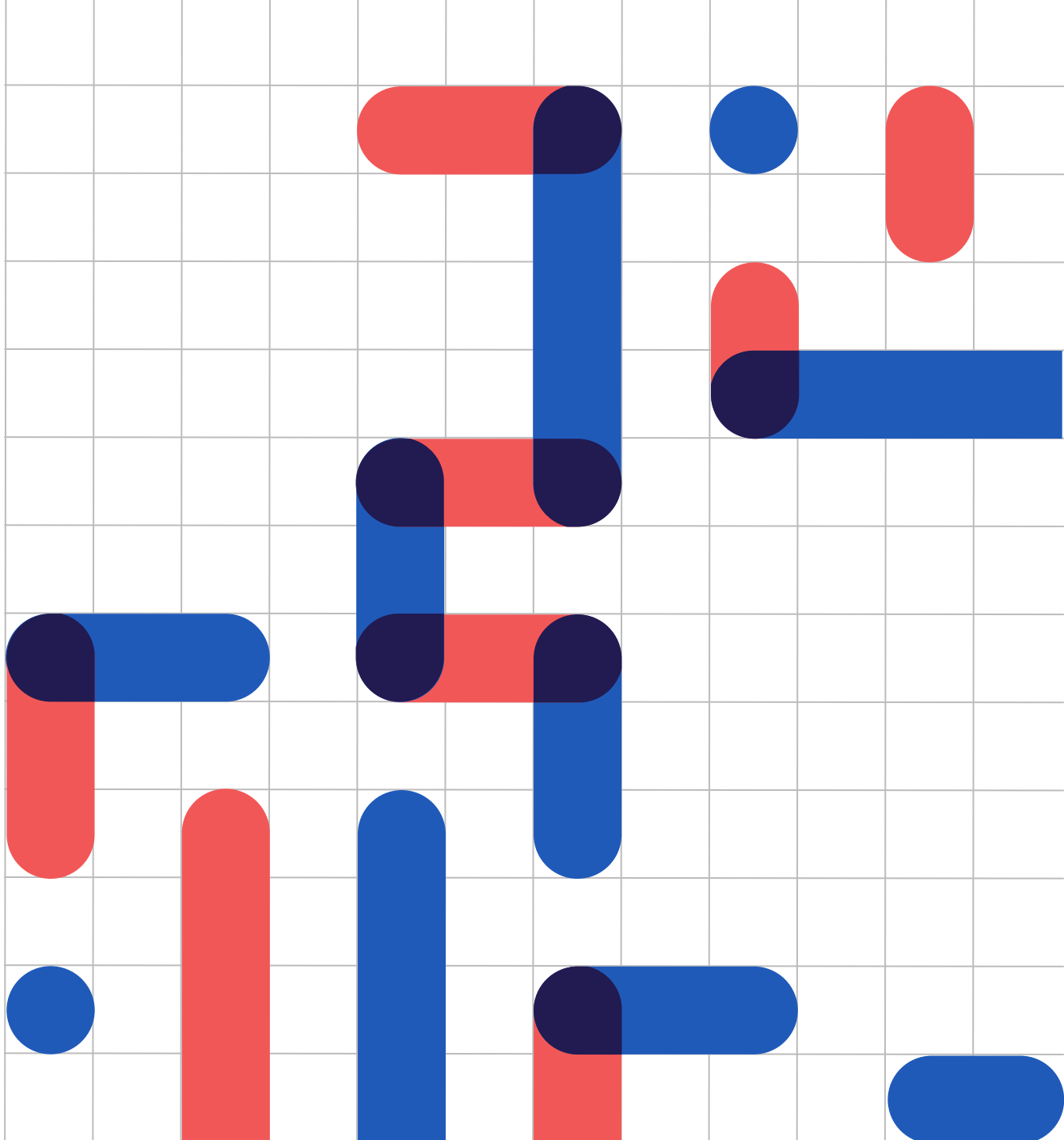


Phables

from fragmented assemblies to bacteriophage genomes

Younginn Park

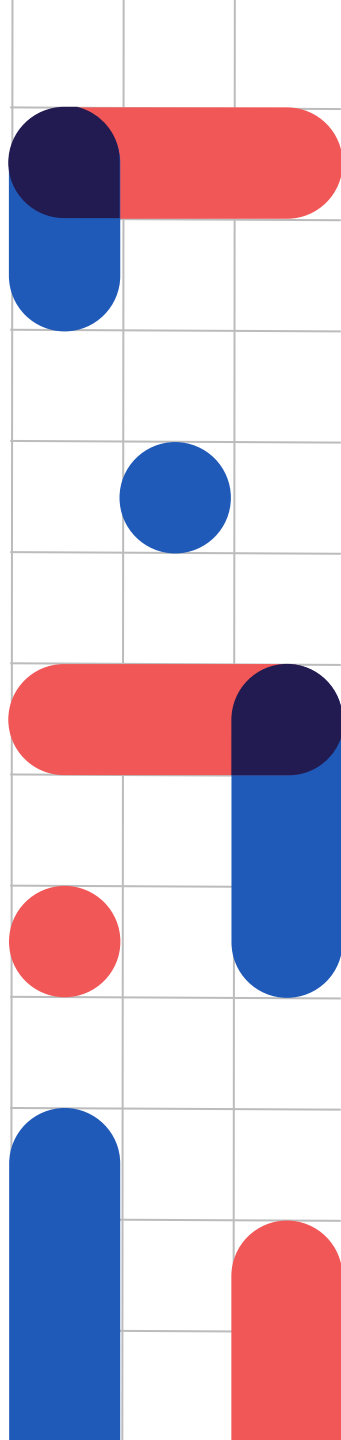


Bacteriophages

Viruses that infect bacteria

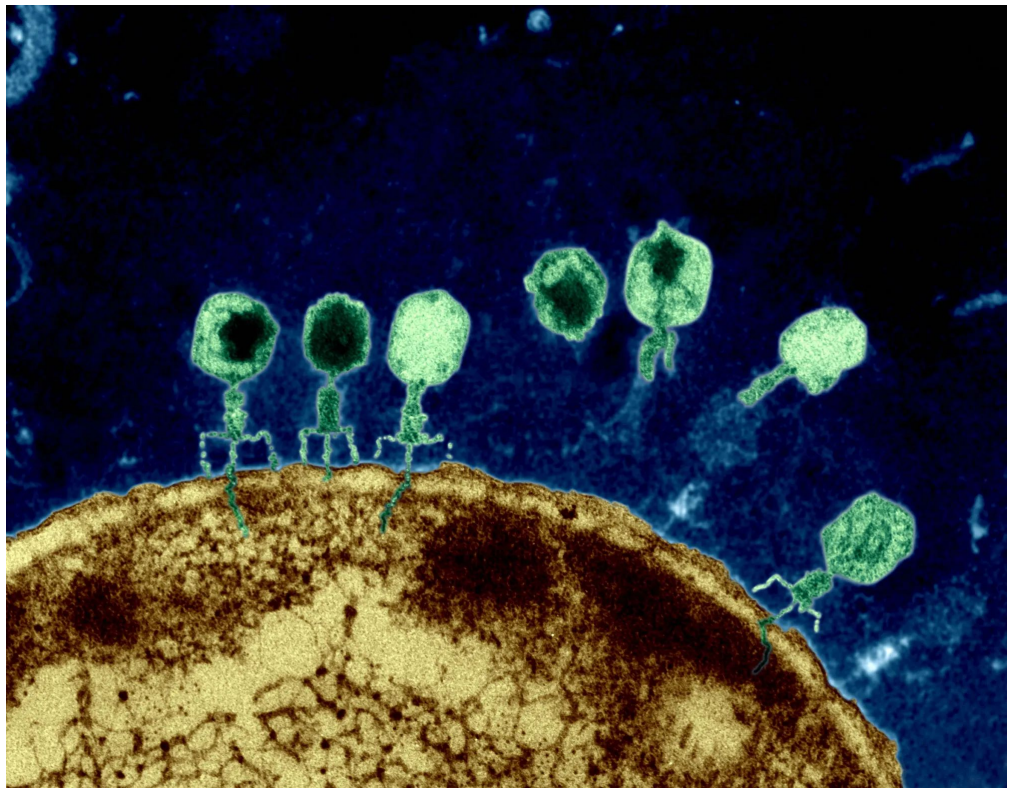
Phages are considered the most abundant biological entity on earth (Comeau et al. 2008)

When sequencing technologies were first developed, phage genomes were the first to be sequenced due to their relatively small genome size (Sanger et al. 1977)



Phages in metagenomes

First metagenomic samples to be sequenced when second-generation sequencing technologies emerged (Breitbart et al. 2002)



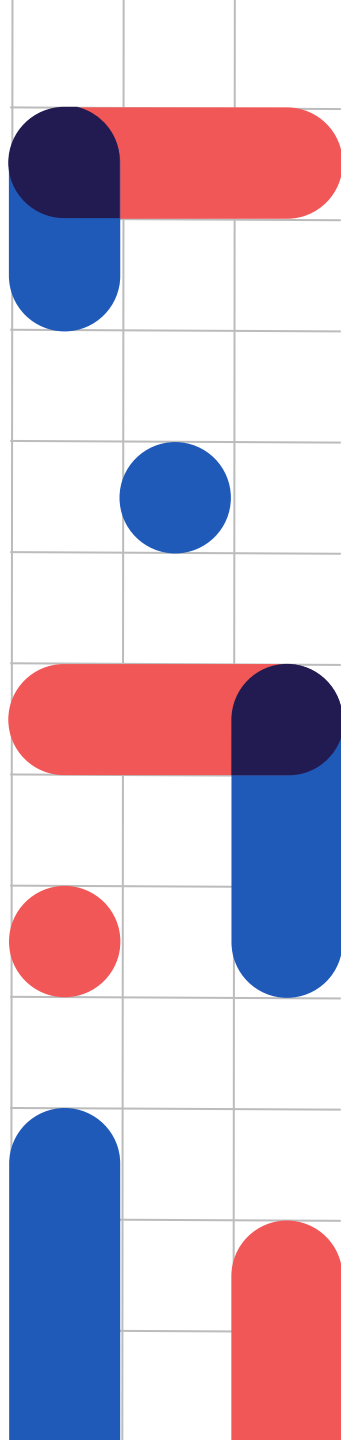
Phages infecting an E. coli cell by injecting DNA through membrane, https://fineartamerica.com/featured/t-bacteriophages-and-e-coli-eye-of-science.html

Why phages?

Medical significance (e.g. inflammatory bowel disease, IBD)

Limited understanding about them

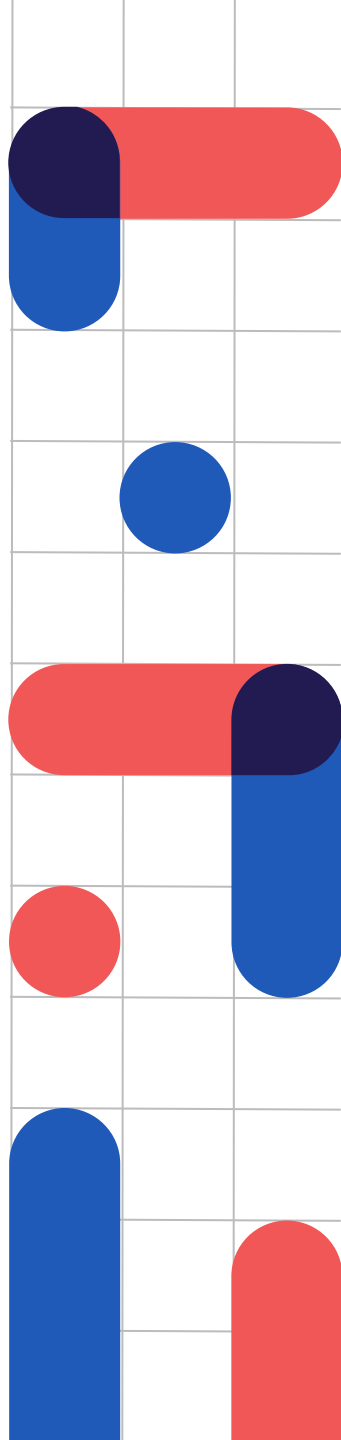
Although countless millions of phage species are thought to exist, only 26 048 complete phage genomes have been sequenced according to the INfrastructure for a PHAge REference Database (INPHARED) (Cook et al. 2021) (as of the September 2023 update).



Challenges

Generating high-quality phage genomes via de novo metagenome assembly is challenging due to the **modular** and **mosaic** nature of phage genomes (Hatfull 2008, Belcaid et al. 2010, Lima-Mendez et al. 2011).

Repeat regions can result in fragmented assemblies and chimeric contigs (Casjens and Gilcrease 2009, Merrill et al. 2016).



Flinders University

Phables

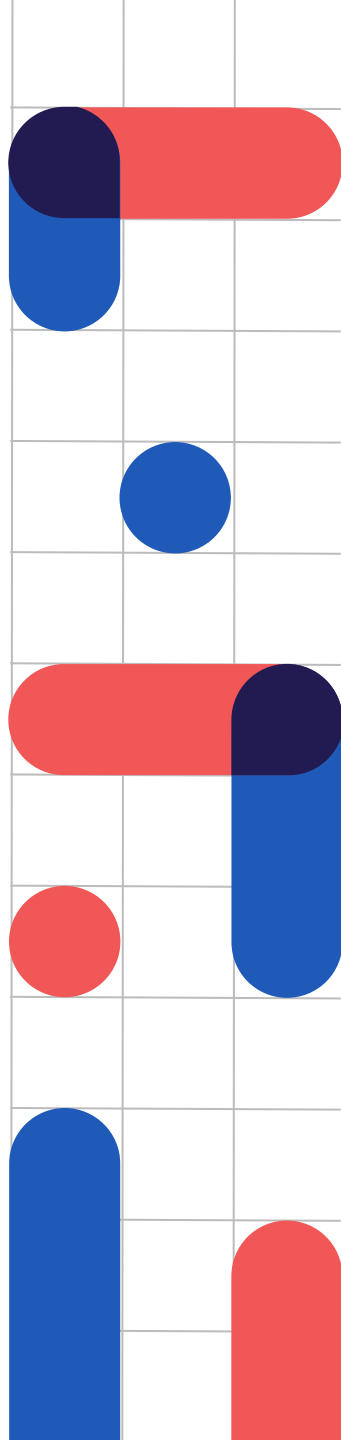
a computational method to resolve phage genomes from fragmented viral metagenome assemblies.

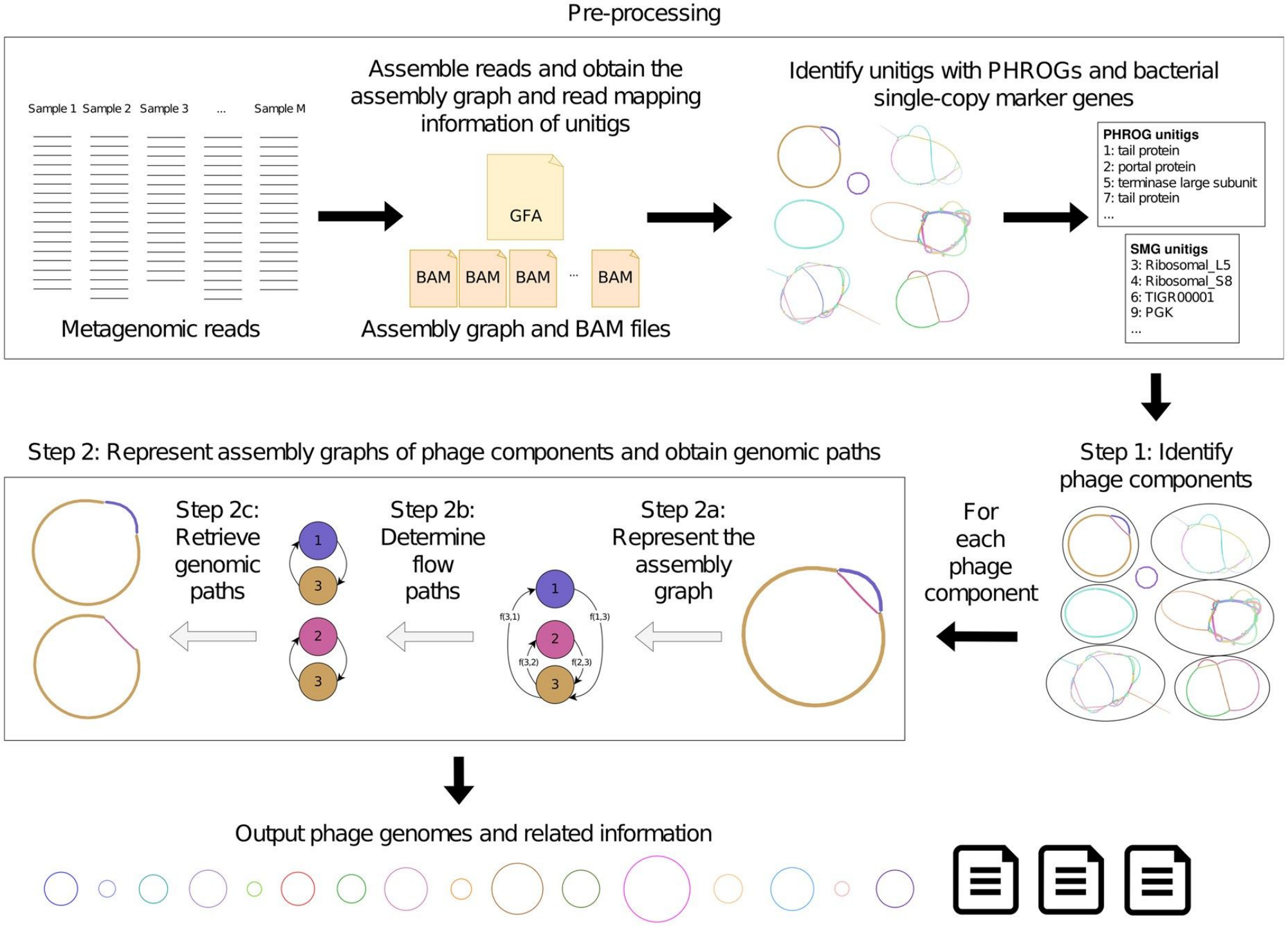
Vijini Mallawaarachchi et al. 2023

Overview

First, Phables identifies phage-like components in the assembly graph using conserved genes.

Second, using read mapping information, graph algorithms and flow decomposition techniques, Phables identifies the most probable combinations of varying phage genome segments within a component, leading to the recovery of accurate phage genome assemblies.




Pre-processing
Sample 1
Sample 2
Sample 3
...
Sample M
Metagenomic reads
Assemble reads and obtain the assembly graph and read mapping information of unitigs
GFA
BAM
BAM
BAM
...
BAM
Assembly graph and BAM files
Identify unitigs with PHROGs and bacterial single-copy marker genes
PHROG unitigs
1: tail protein
2: portal protein
5: terminase large subunit
7: tail protein
...
SMG unitigs
3: Ribosomal_L5
4: Ribosomal_S8
6: TIGR00001
9: PGK
...
Step 1: Identify phage components
For each phage component
Step 2: Represent assembly graphs of phage components and obtain genomic paths
Step 2a: Represent the assembly graph
Step 2b: Determine flow paths
Step 2c: Retrieve genomic paths
f(3,1)
f(1,3)
f(3,2)
f(2,3)
Output phage genomes and related information

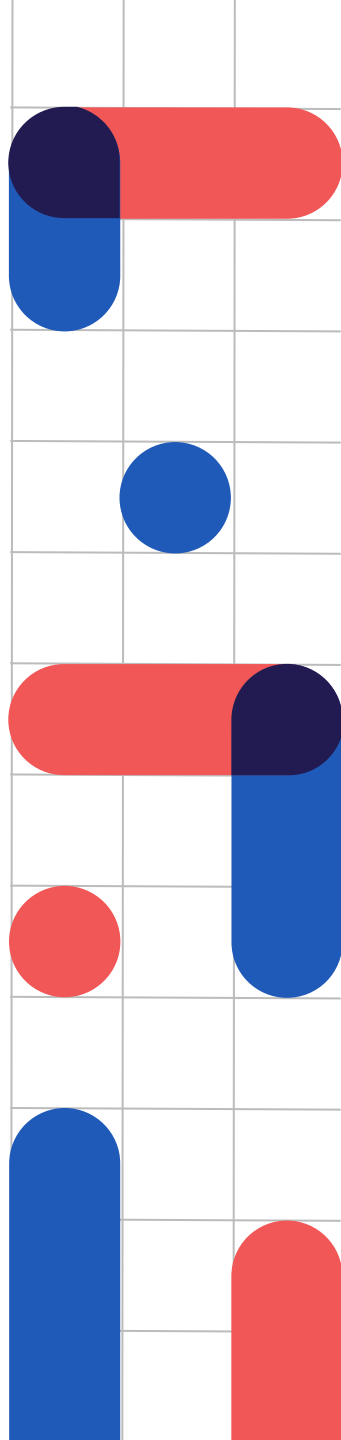
Preprocessing

Use assembly tool like Hecatomb (Roach et al. 2022a) to generate an assembly graph (GFA format).

The unitig sequences are extracted from the assembly graph, and the raw sequencing reads are mapped to the unitigs using Minimap2 (Li 2018) and Samtools (Li et al. 2009)

PHROGs

PHROGs are viral protein families commonly used to annotate prokaryotic viral sequences. MMSeqs2 (Steinegger and Söding 2017) is used to identify PHROGs (Prokaryotic Virus Remote Homologous Groups) in unitigs using an identity cutoff of 30% and an e-value of less than (by default).

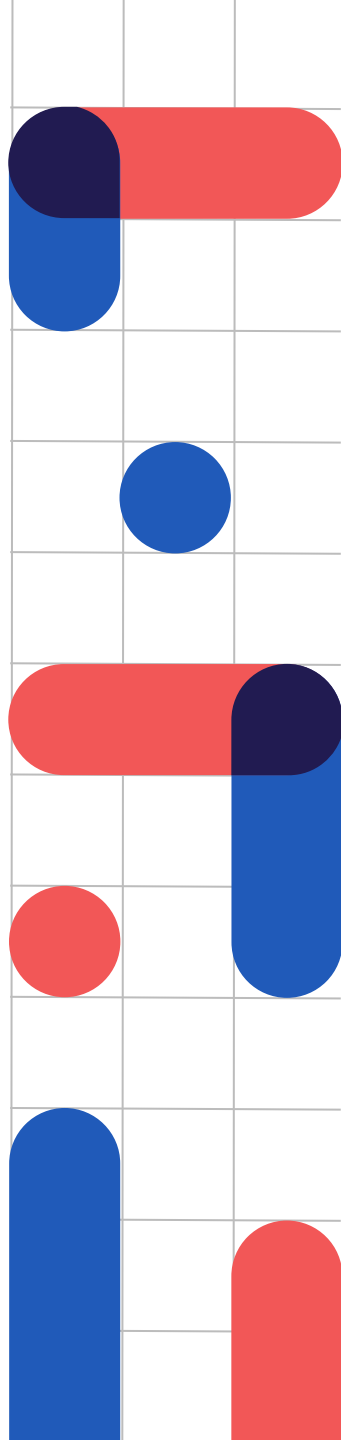


SMGs

Next, Phables identifies unitigs containing bacterial single-copy marker genes (SMGs).

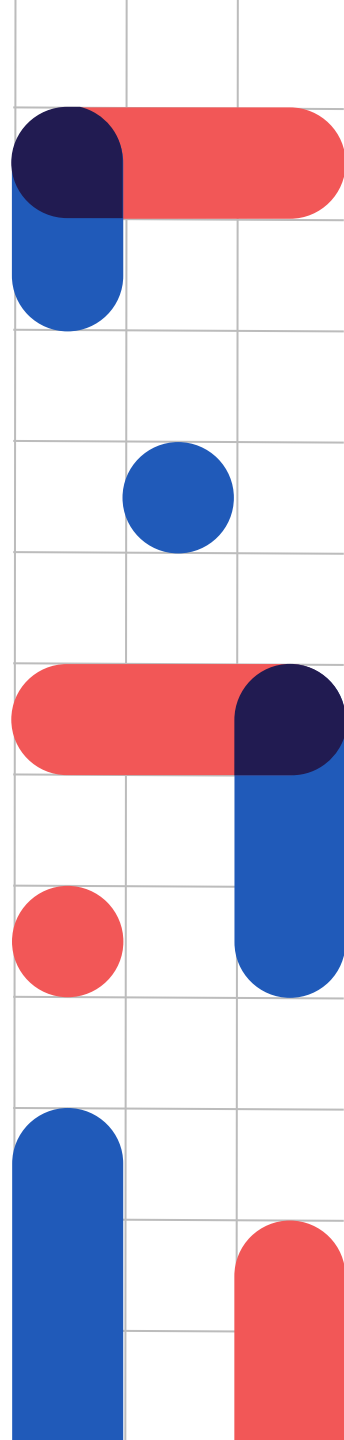
Most bacterial genomes have conserved genes known as single-copy marker genes (SMGs) that appear only once in a genome (Dupont et al. 2012, Albertsen et al. 2013).

FragGeneScan and HMMER used to identify sequences with SMGs



SMGs

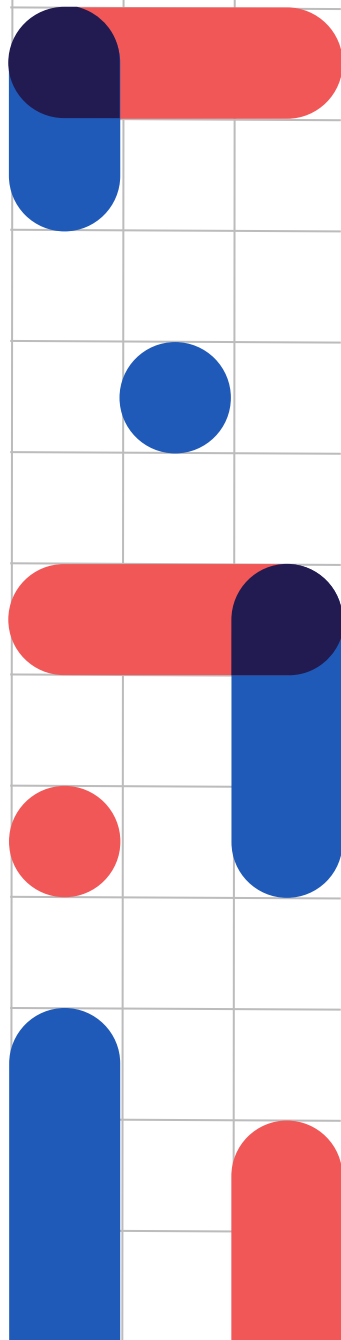
Phables identifies components from the final assembly graph where all of its unitigs do not have any bacterial SMGs (identified from the preprocessing step). This ensures that the components are not prophages.



Graph definition

Graph of a phage component G = <V,E>, where V vertices are unitig sequences that make up a phage component and directed edges (VxV) with connections between unitigs

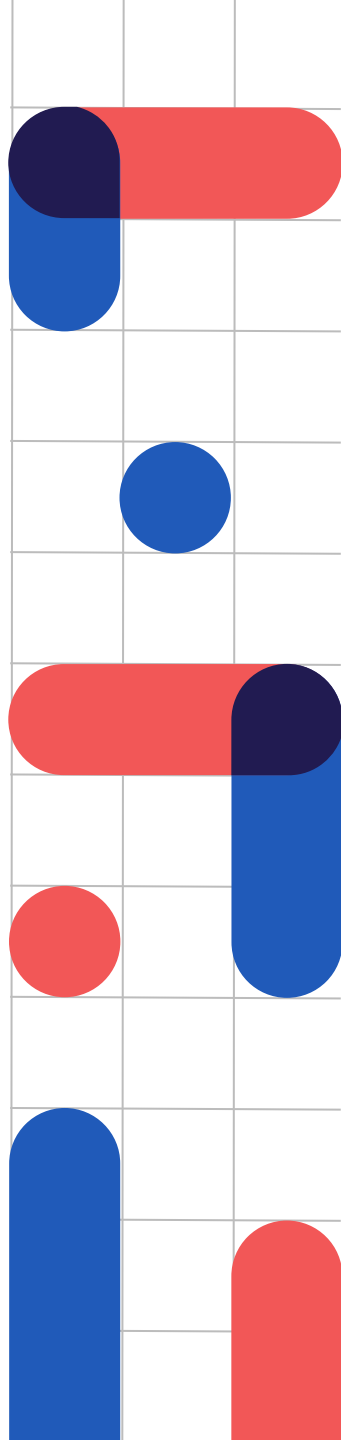
The weight of each edge (w_e(u->v)) is set to the minimum of the read coverage values of the two unitigs u and v

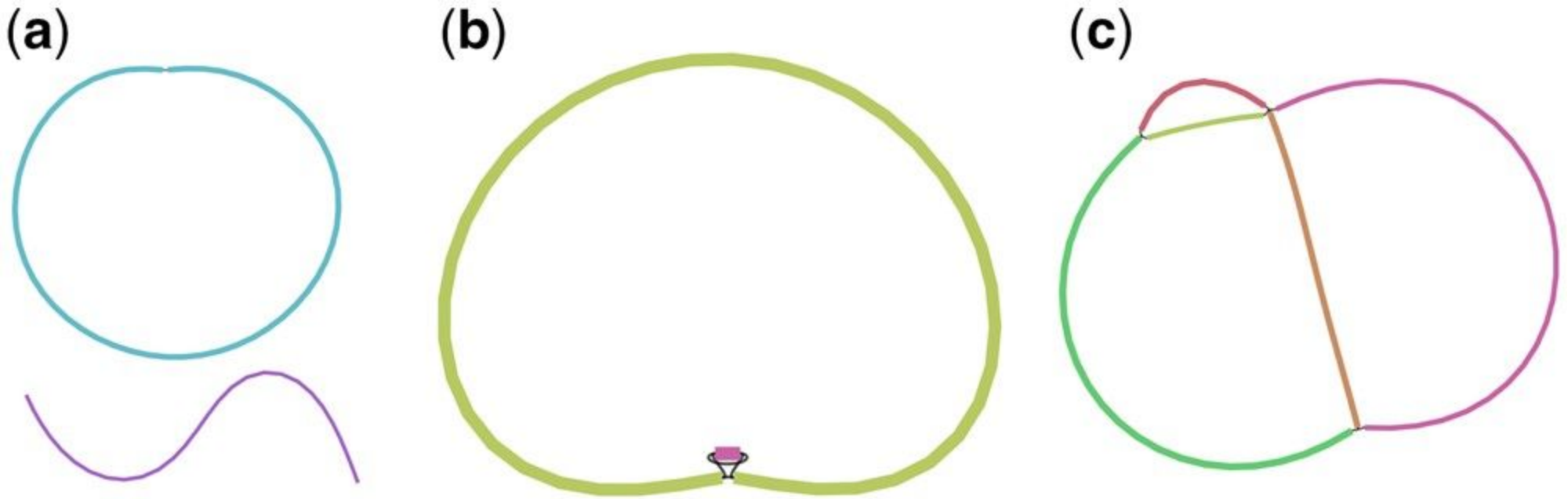


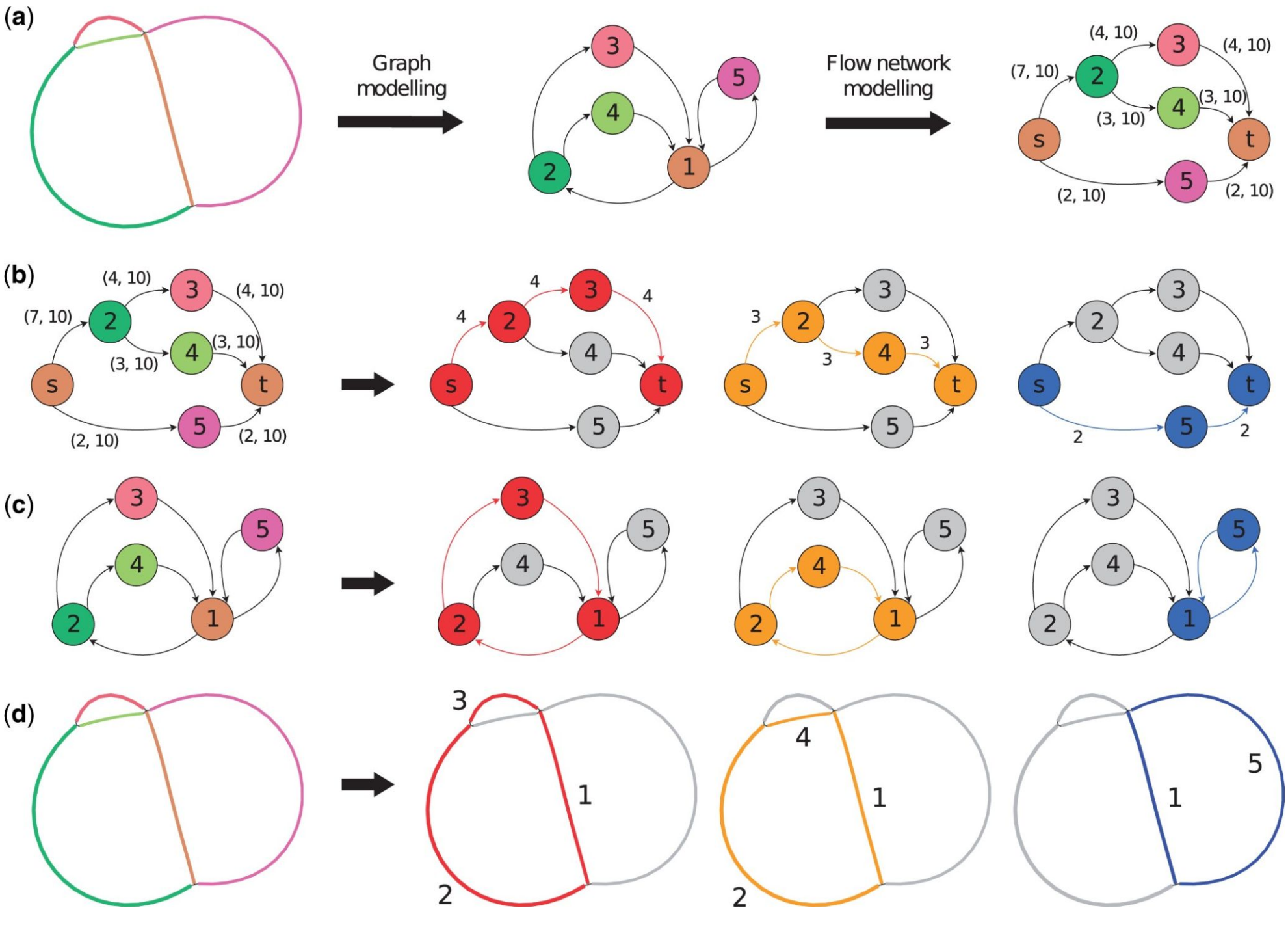
Flow decomposition

Breaking down the total flow in a network into simpler components, such as paths or cycles, while satisfying certain constraints or optimizing specific objectives.

Obtaining genomic paths by flow decomposition in the graphs.

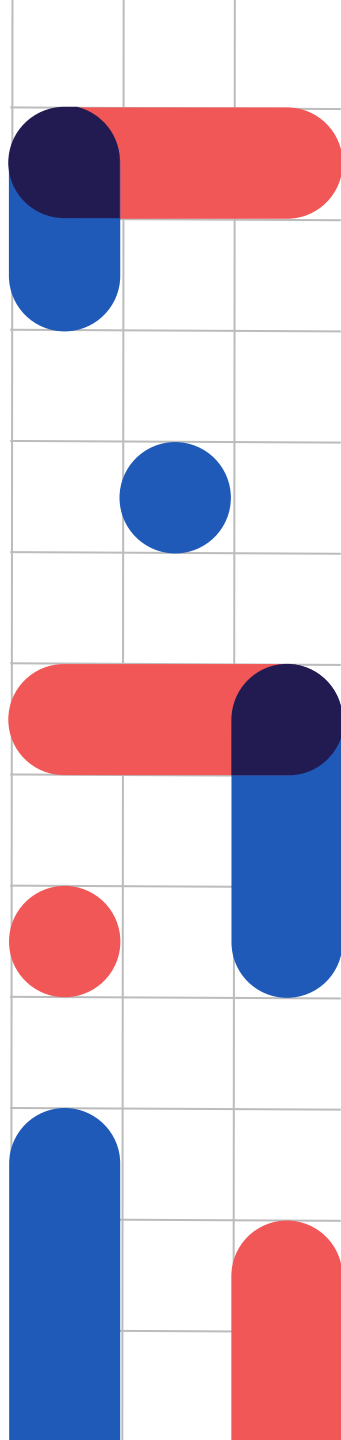



(a)
(b)
(c)


(a)
Graph modelling
Flow network modelling
(4, 10)
(7, 10)
(3, 10)
(2, 10)
(b)
(c)
(d)

Benchmarks

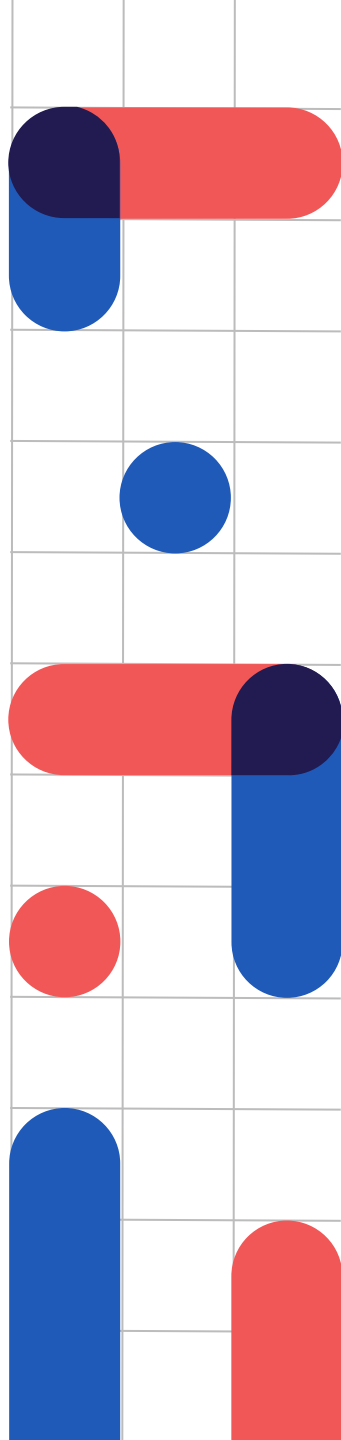
- Simulated dataset (simPhage)
- NCBI viral genomes against PHAMB



simPhage dataset

A simulated phage dataset (referred to as ‘simPhage’) to evaluate Phables.

- Enterobacteria phage P22 (AB426868)
- Enterobacteria phage T7 (NC_001604)
- Staphylococcus phage SAP13 TA-2022 (ON911718)
- Staphylococcus phage SAP2 TA-2022 (ON911715)



simPhage

Table 1. Evaluation results for the genomes resolved from Phables for the simPhage dataset.

Genome	Simulated coverage	Phables predicted coverage	Genome coverage (%)
P22	100	100	99.947
T7	150	150	99.599
SAP13 TA-2022	200	206	100.00
SAP2 TA-2022	400	401	92.406

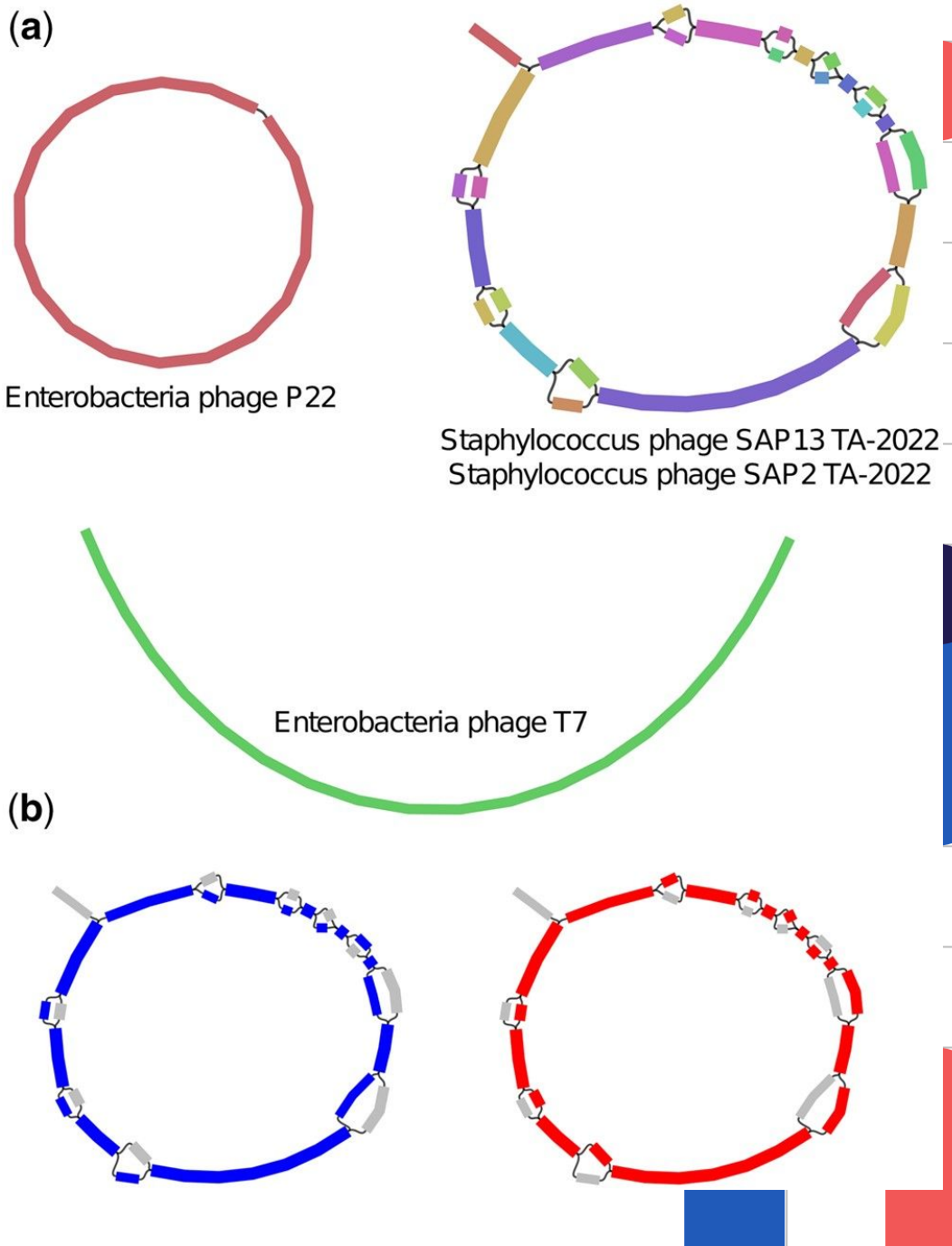
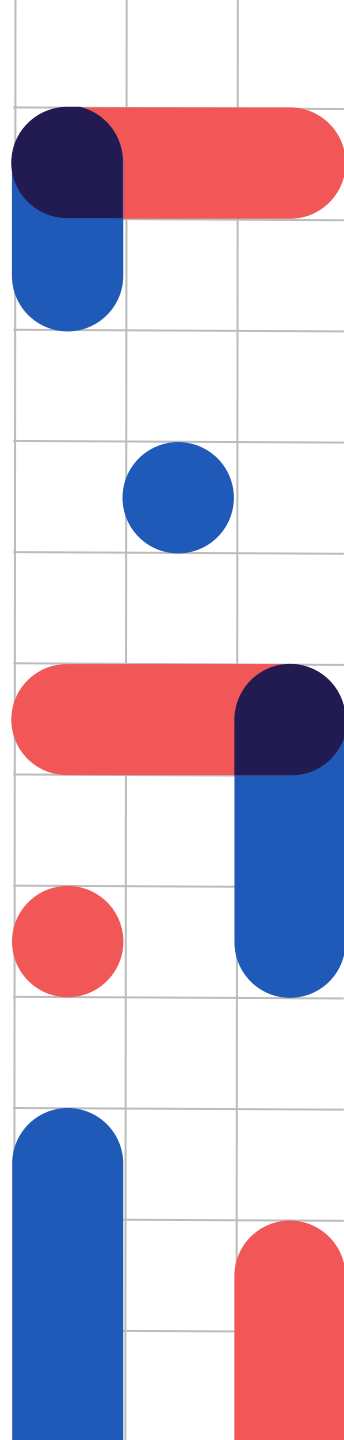


Figure (left). simPhage assembly graph. Visualization of the (a) assembly graph from the simPhage dataset with phage components and (b) resolution of two paths (red and blue) from the Staphylococcus phage component.

NCBI viral genomes

Four metagenomic datasets from NCBI

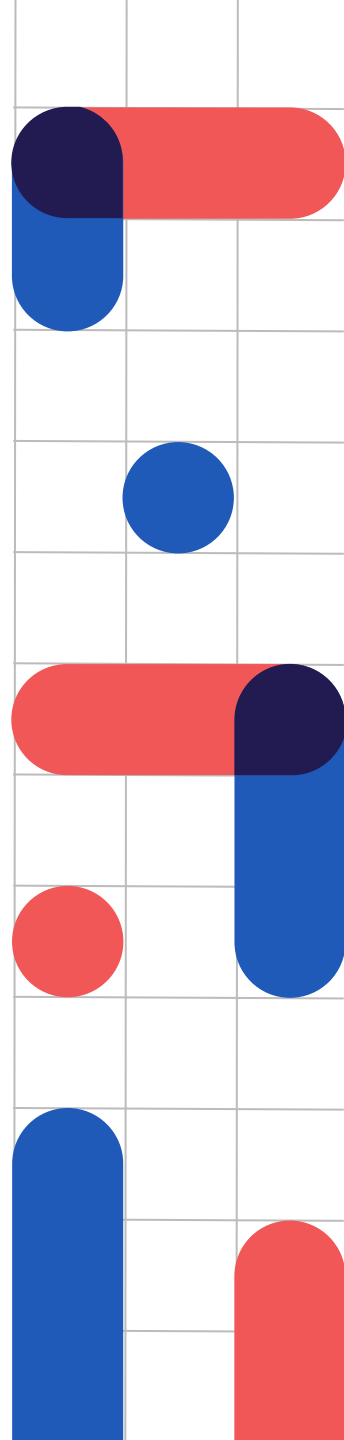
- Lake Water
- Paddy Soil
- Wastewater
- Stool samples from patients with IBD (inflammatory bowel disease)



PHAMB

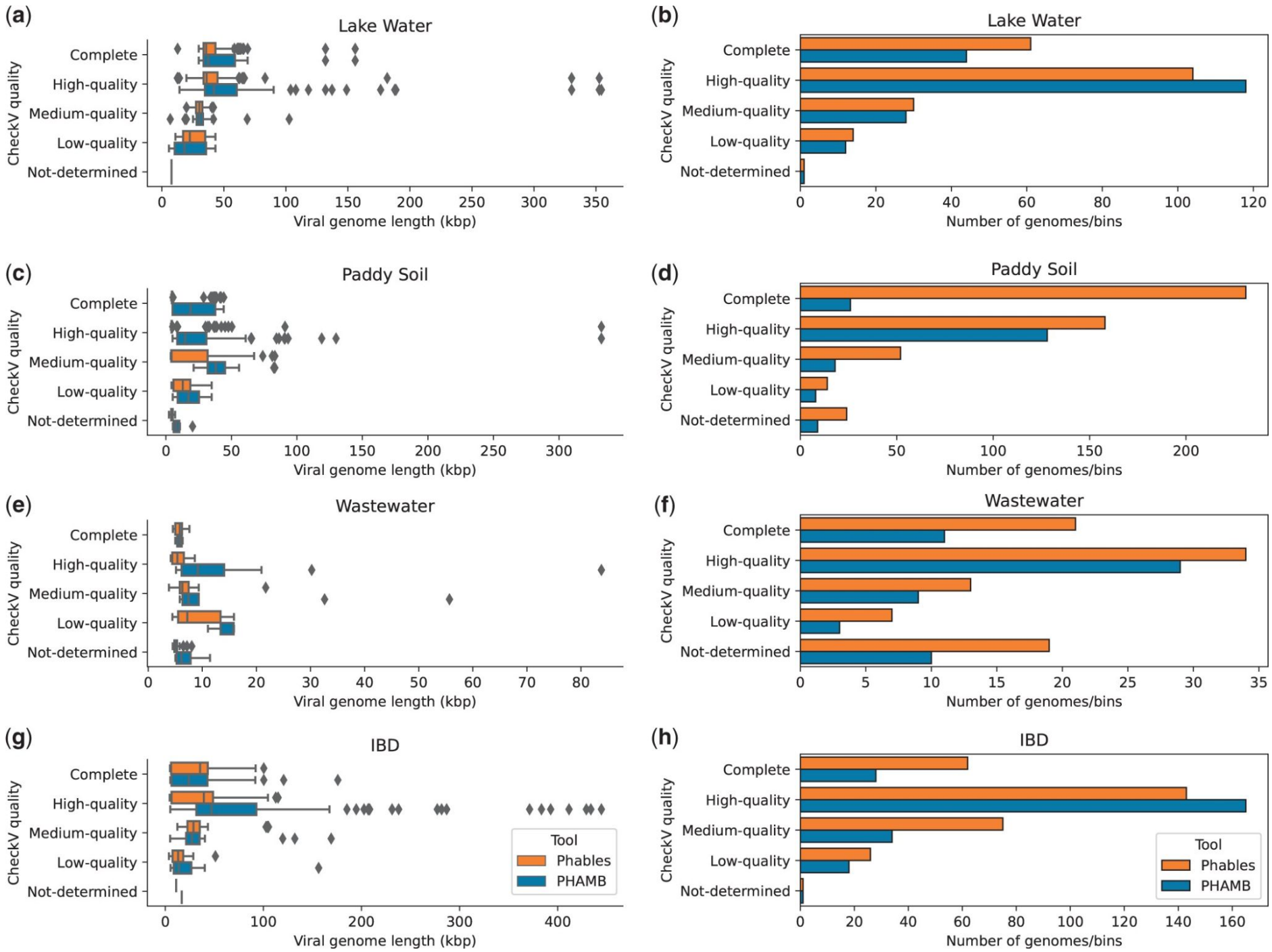
Benchmarked against PHAMB (Johansen et al. 2022) (version 1.0.1), a viral identification tool that predicts whether MAGs represent phages and outputs genome sequences.

PHAMB takes binning results from a metagenomic binning tool and predicts bins that contain bacteriophage sequences.



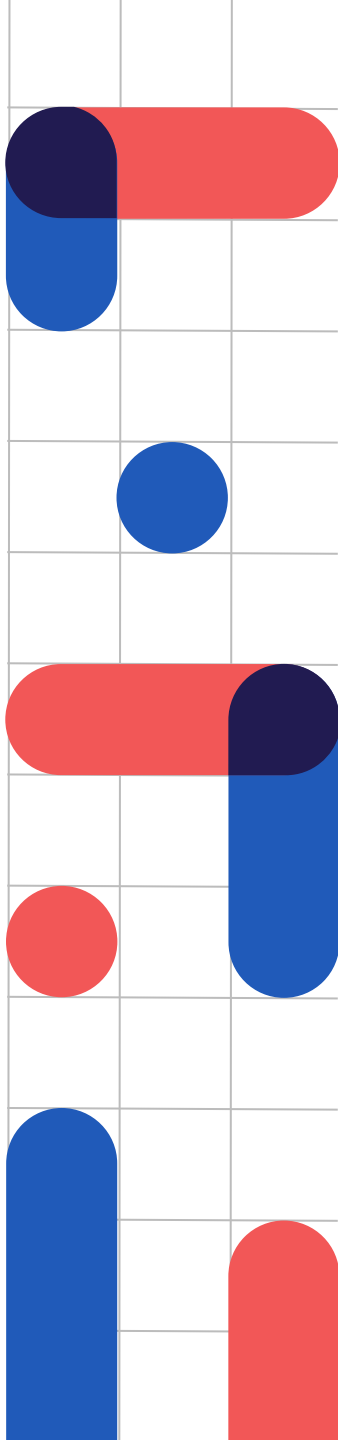
Evaluation criteria

CheckV version 1.0.1 (Nayfach et al. 2021a) used for genomes evaluation - it compares bins/genomes against a large database of complete viral genomes by completeness, quality.


(a)
Lake Water
CheckV quality
Complete
High-quality
Medium-quality
Low-quality
Not-determined
0 50 100 150 200 250 300 350
Viral genome length (kbp)
(b)
Lake Water
0 20 40 60 80 100 120
Number of genomes/bins
(c)
Paddy Soil
0 50 100 150 200 250 300
Viral genome length (kbp)
(d)
Paddy Soil
0 50 100 150 200
Number of genomes/bins
(e)
Wastewater
0 10 20 30 40 50 60 70 80
Viral genome length (kbp)
(f)
Wastewater
0 5 10 15 20 25 30 35
Number of genomes/bins
(g)
IBD
0 100 200 300 400
Viral genome length (kbp)
Tool
Phables
PHAMB
(h)
IBD
0 20 40 60 80 100 120 140 160
Number of genomes/bins

Discussion

- Further work needed for metagenomes of mixed-microbial communities
- Support for long-read assemblies
- Extend the capabilities to recovering high-quality eukaryotic viral genomes from metagenomes



References

Vijini Mallawaarachchi, Michael J Roach, Przemyslaw Decewicz, Bhavya Papudeshi, Sarah K Giles, Susanna R Grigson, George Bouras, Ryan D Hesse, Laura K Inglis, Abbey L K Hutton, Elizabeth A Dinsdale, Robert A Edwards, *Phables: from fragmented assemblies to high-quality bacteriophage genomes*, Bioinformatics, Volume 39, Issue 10, October 2023, btad586, https://doi.org/10.1093/bioinformatics/btad586