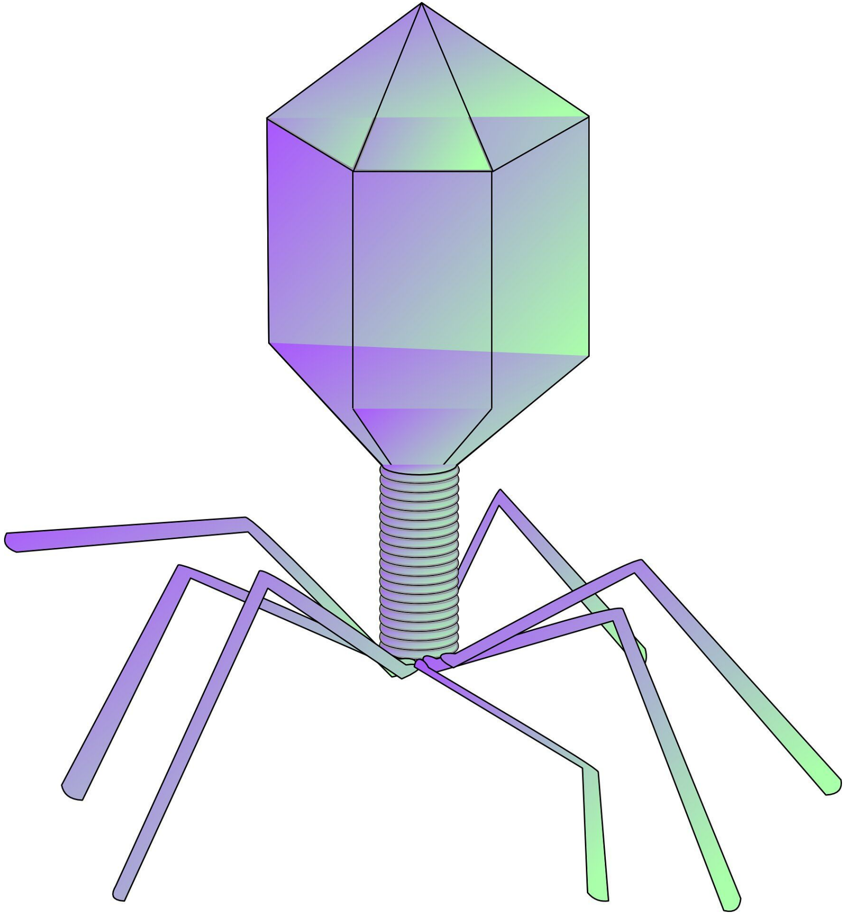


Bacteriophage Identification in Metagenomes

Polina Tikhonova
DAWG, SP25

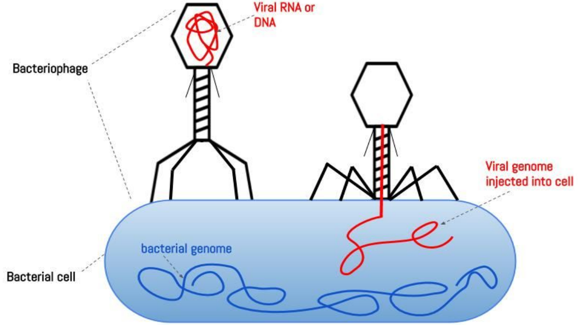
Bacteriophage (phage)

IS a virus that infects and replicates within bacteria and archaea.

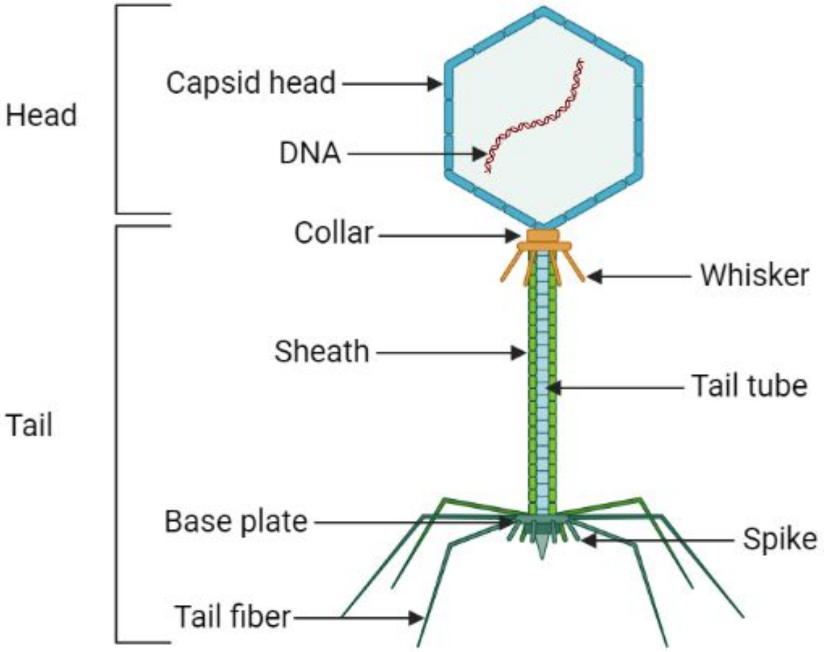
IT'S GENETIC MATERIAL COULD BE

- DNA or RNA
- Single or Double stranded

BACTERIA INFECTION PROCESS

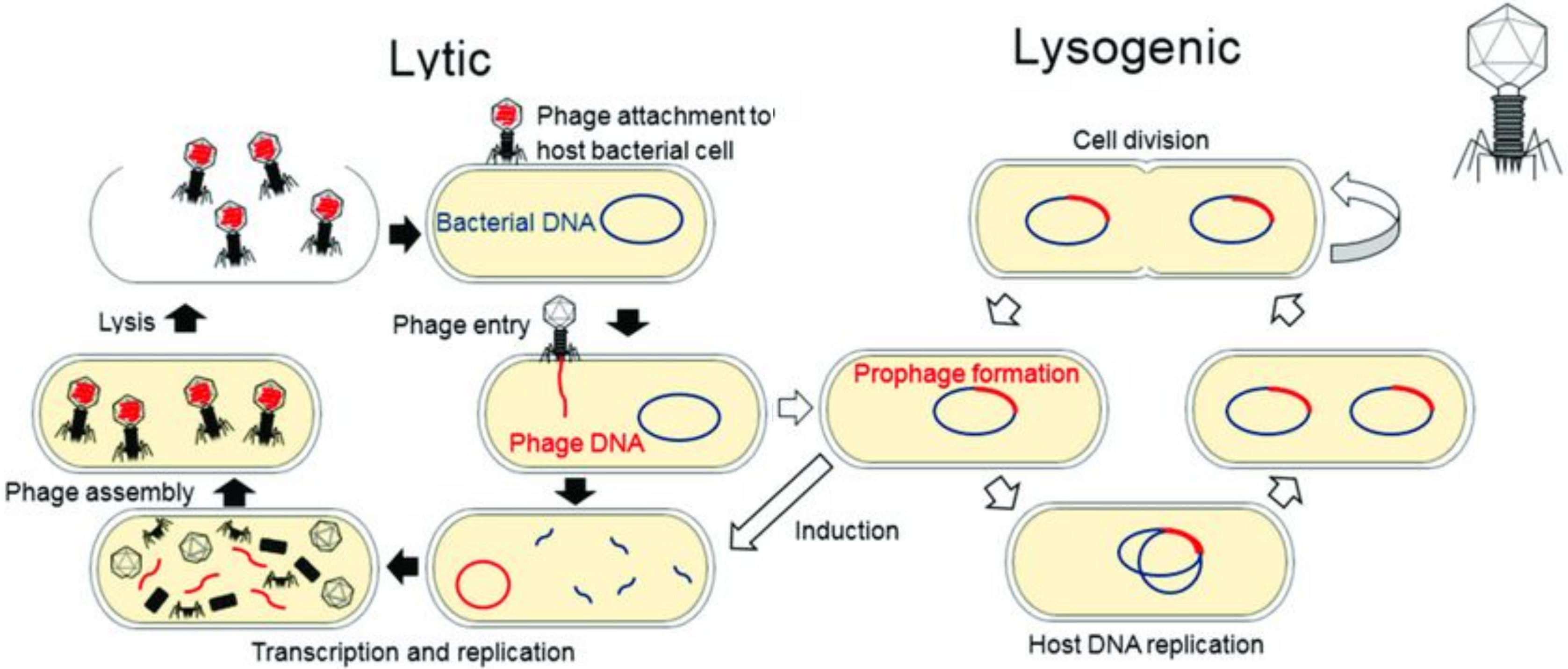


Source:https://www.whatisbiotechnology.org/index.php/science/summary/phage-therapy/phage-therapy-uses-viruses-that-attack-bacteria-to-treat



Source:https://www.ahmadcoaching.com/2021/02/life-cycle-of-bacteriophage-lytic-lysogenic-induction.html

Reproduction cycle of Bacteriophage



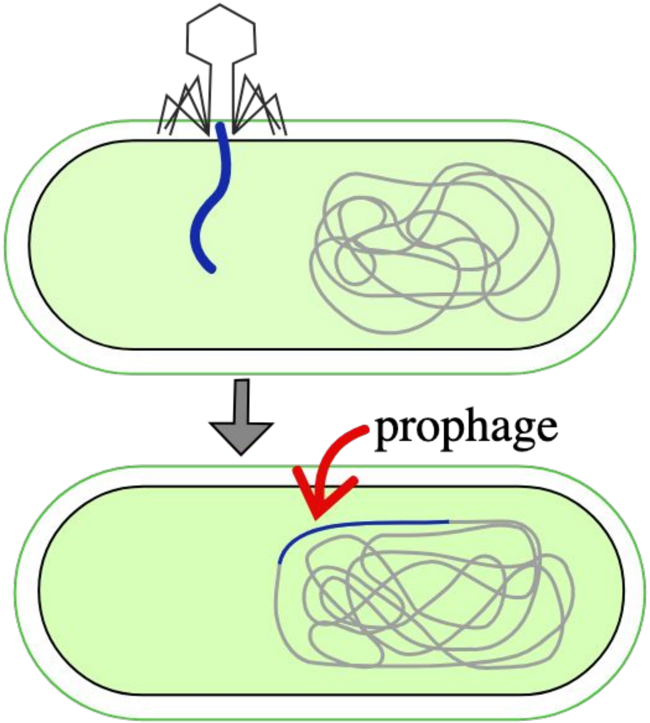
DOI: 10.3390/pathogens8030100

Prophage

is a bacteriophage genetic material incorporated in host DNA

Prophages have a great influence on bacterial evolution.

- They are a major source of horizontal gene transfer in bacteria.
- Prophages could increase bacteria pathogenicity by carrying toxin genes.
- Through HGT, prophages also could make bacteria antibiotic-resistant.



Why to study Bacteriophages or Prophages?

Fun facts

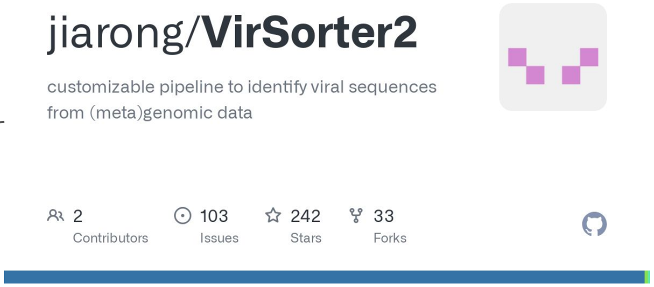
- Prophages are the most abundant biological form in the biosphere with an estimated 10^31 phages (more diverse than bacteria).
- Water surface of Earth harbours large portion of prophage diversity.
- E.coli-targeting phages seem to dominate
- Prophages could also target commensal bacteria (e.g. *Bacteroides fragilis*)

- To understand the mechanisms of arising abx-resistance among bacteria
- To identify bacterial pathogenicity sources
- To build phage-therapy cocktails against bacterial infections
- Delivery of modification CRISPR-Cas9 systems to the bacteria

How could we identify prophages in bacteria genomes?

Proksee

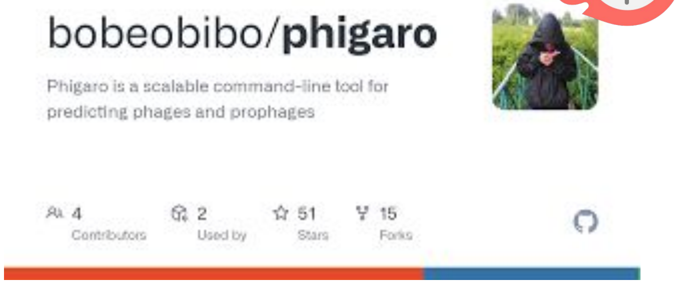
PHASTEST



jiarong/**VirSorter2**

customizable pipeline to identify viral sequences from (meta)genomic data

2 Contributors 103 Issues 242 Stars 33 Forks



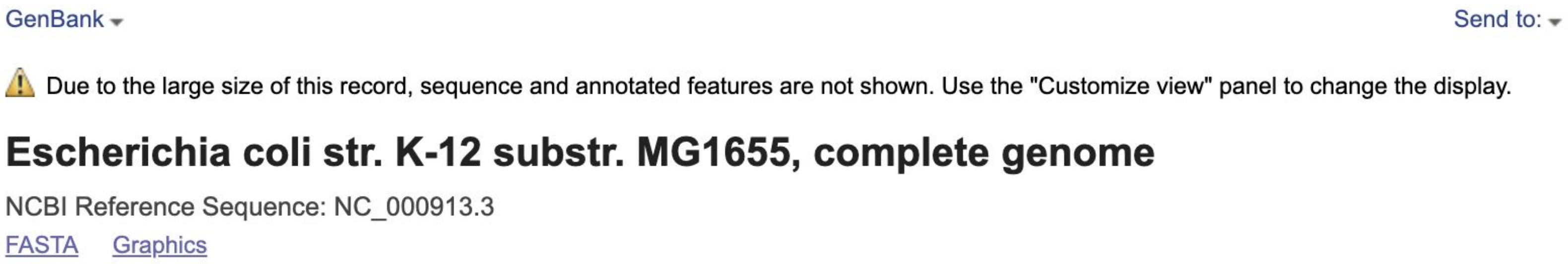
bobeobibo/**phigaro**

Phigaro is a scalable command-line tool for predicting phages and prophages

4 Contributors 2 Used by 51 Stars 15 Forks

Let’s start looking at E.coli in Proksee

https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3



⚠ Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

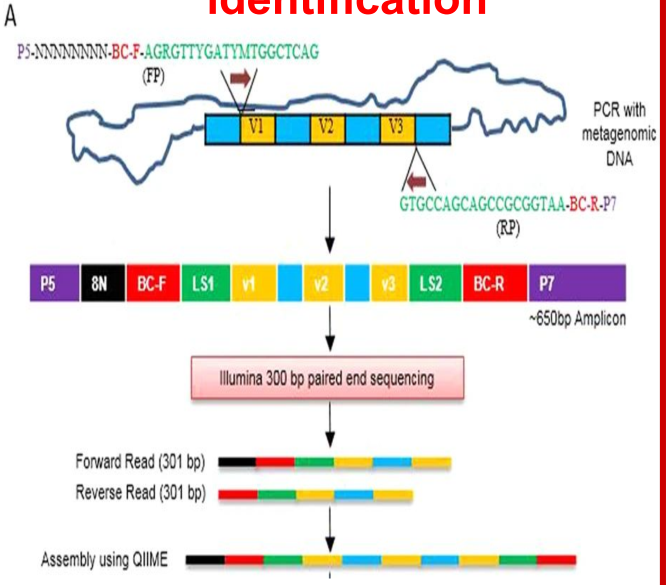
Escherichia coli str. K-12 substr. MG1655, complete genome

NCBI Reference Sequence: NC_000913.3

How to look for prophages in our microbial data?

16s RNA

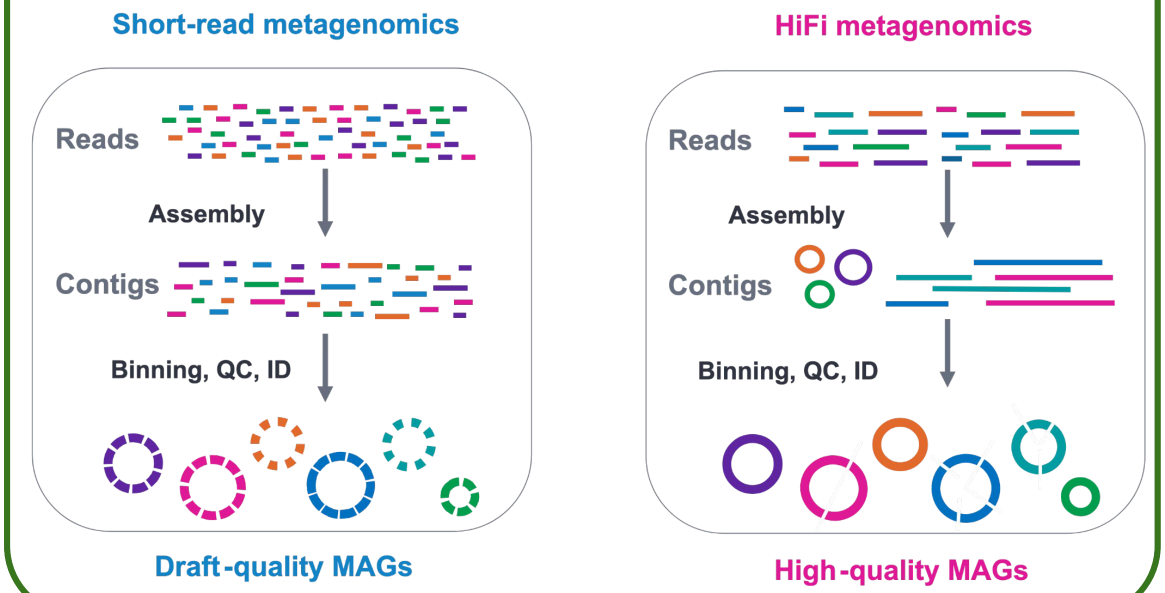
❌ **Not suitable for prophage identification**



DOI:10.1038/srep25882

Metagenomics sequences

✓ **Good for prophage identification**



Short-read metagenomics

Reads

Assembly

Contigs

Binning, QC, ID

Draft-quality MAGs

HiFi metagenomics

Reads

Assembly

Contigs

Binning, QC, ID

High-quality MAGs

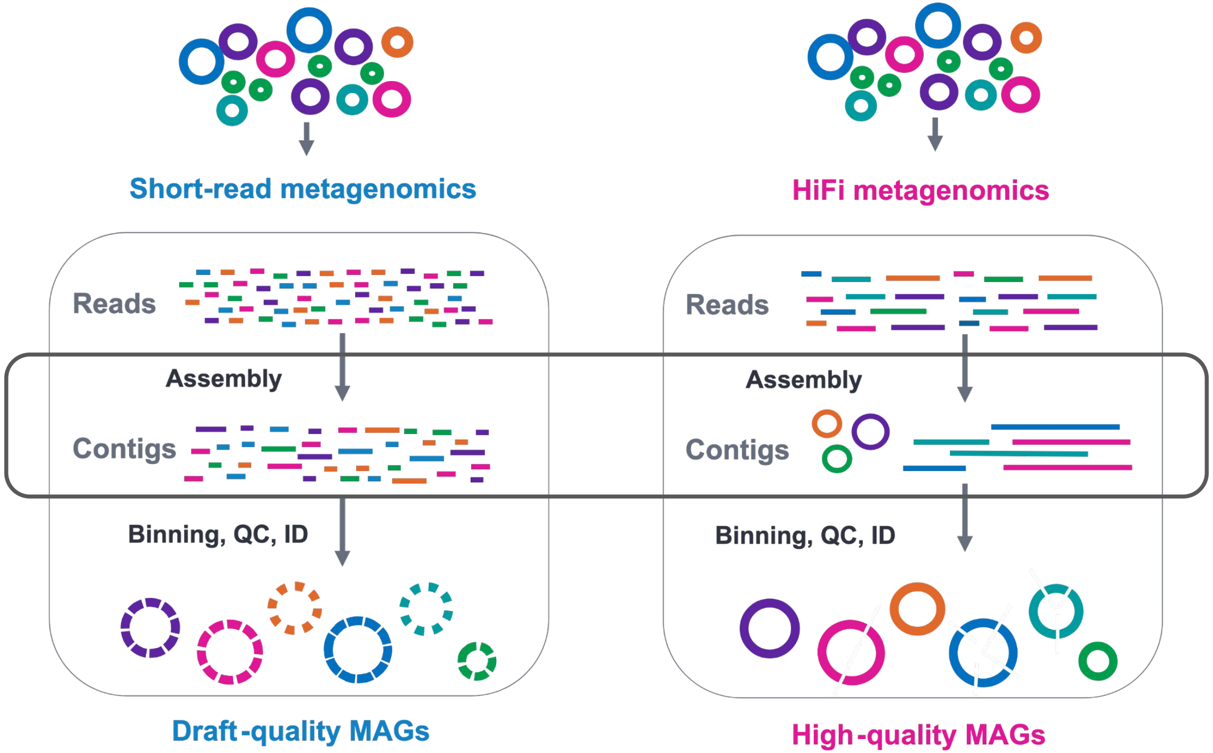
Source: https://www.pacb.com/blog/sequencing-101-metagenome-assembled-genomes/

How to look for prophages in our microbial data?

Step 1.
Use metagenomics sequences.

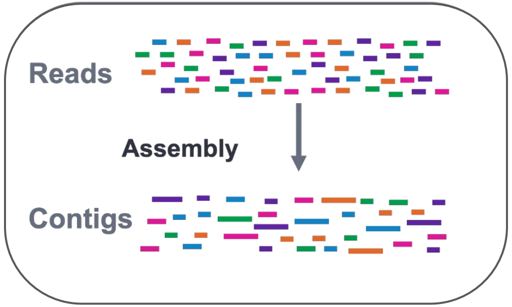
Step 2.
Assemble reads into contigs.

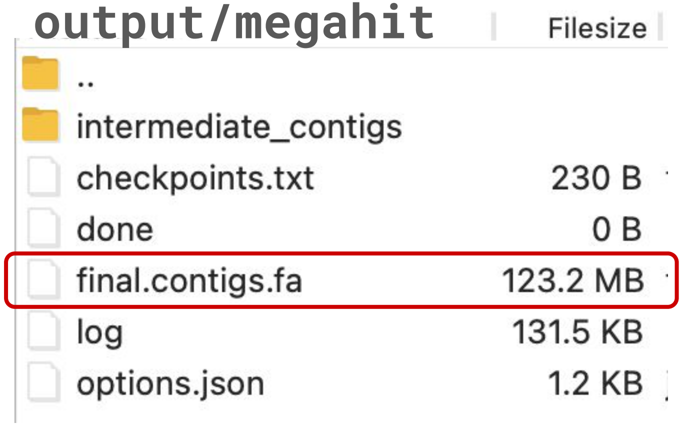
Step 3.
Find prophage regions.



Step 2. Assemble reads into contigs

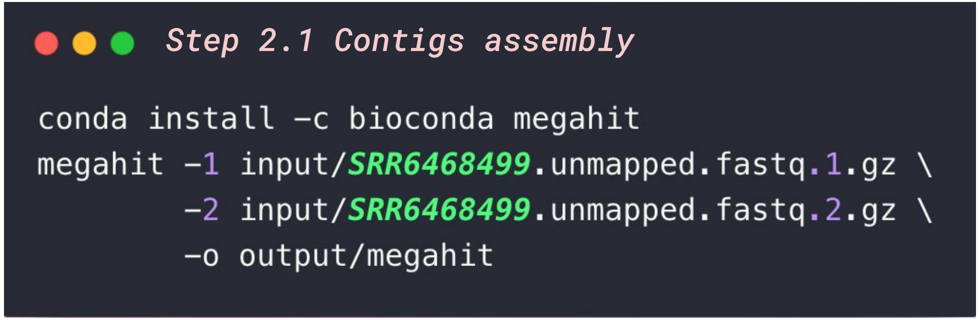
Short-read metagenomics





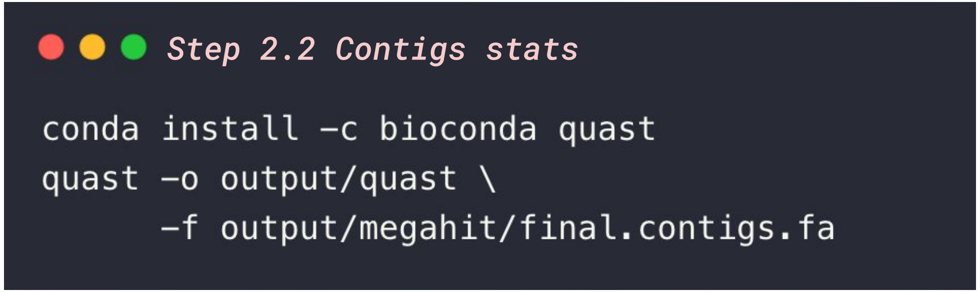
output/megahit

	Filesize
..	
intermediate_contigs	
checkpoints.txt	230 B
done	0 B
final.contigs.fa	123.2 MB
log	131.5 KB
options.json	1.2 KB



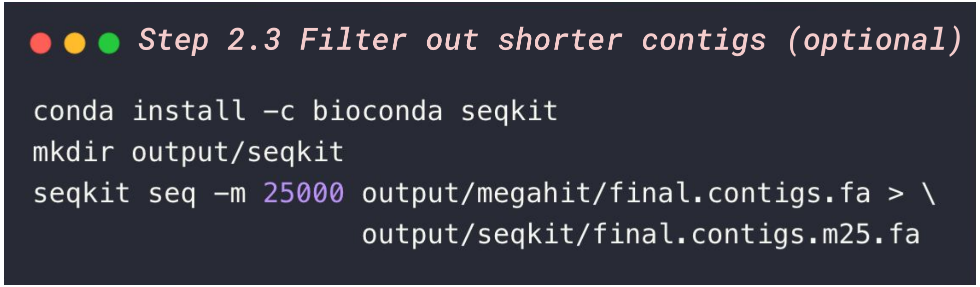
Step 2.1 Contigs assembly

```
conda install -c bioconda megahit
megahit -1 input/SRR6468499.unmapped.fastq.1.gz \
        -2 input/SRR6468499.unmapped.fastq.2.gz \
        -o output/megahit
```



Step 2.2 Contigs stats

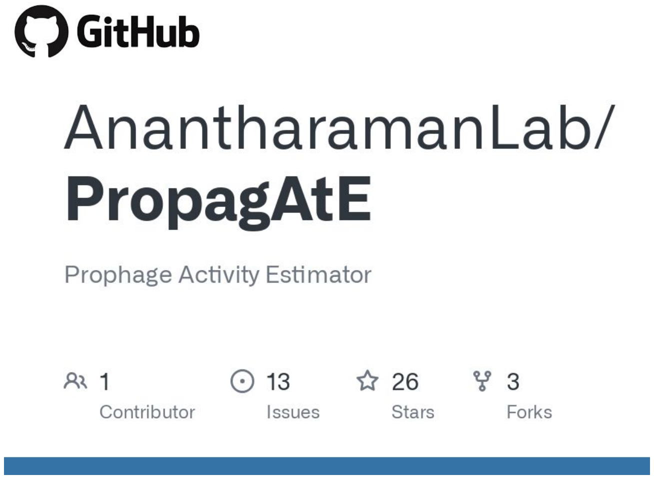
```
conda install -c bioconda quast
quast -o output/quast \
      -f output/megahit/final.contigs.fa
```



Step 2.3 Filter out shorter contigs (optional)

```
conda install -c bioconda seqkit
mkdir output/seqkit
seqkit seq -m 25000 output/megahit/final.contigs.fa > \
                    output/seqkit/final.contigs.m25.fa
```

Let’s see if the identified prophages are active?!



GitHub

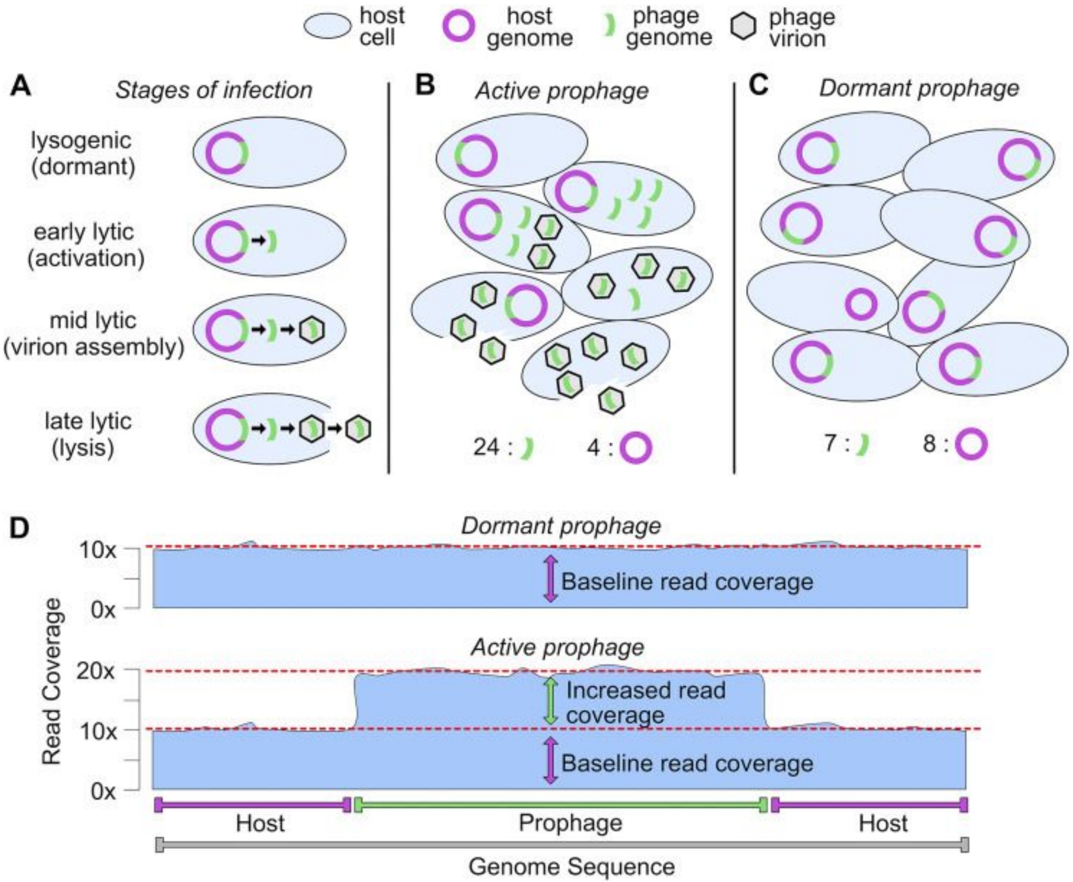
AnantharamanLab/
PropagAtE

Prophage Activity Estimator

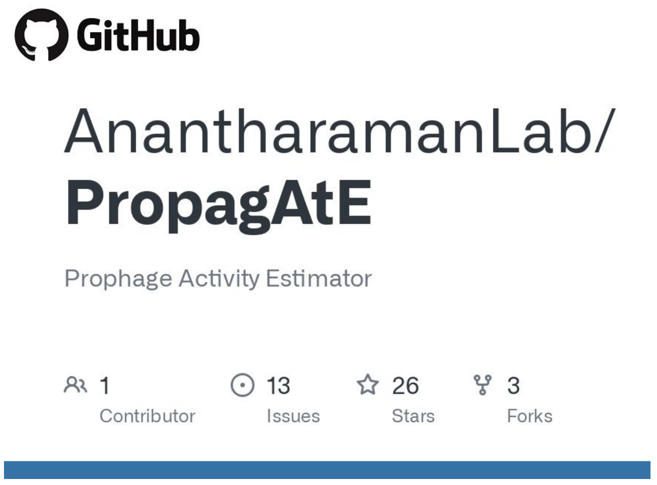
1	13	26	3
Contributor	Issues	Stars	Forks

input/

File 1. genome.fasta
File 2. prophage_coordinates.csv
File 3. reads.fastq.1.gz
reads. fastq.2.gz



https://doi.org/10.1128/msystems.00084-22



GitHub

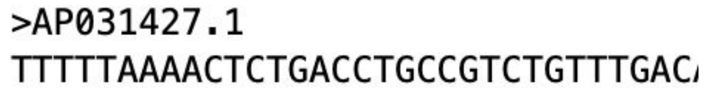
AnantharamanLab/ **PropagAtE**

Prophage Activity Estimator

1 Contributor · 13 Issues · 26 Stars · 3 Forks

Fasta ID should be identical to scaffold ID!!!

input/genome.fasta



```
>AP031427.1
TTTTTAAAACTCTGACCTGCCGTCTGTTTGAC
```

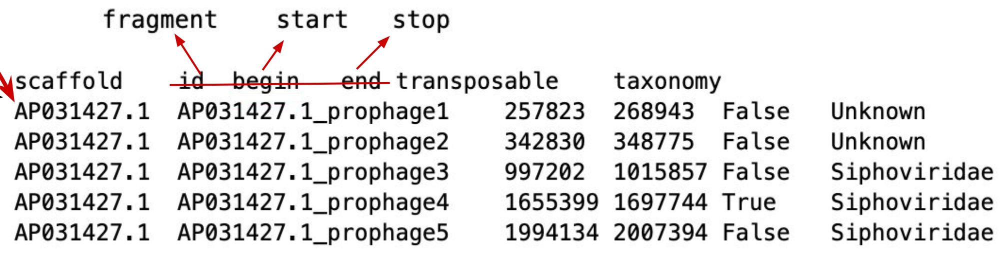
input/

File 1. genome.fasta

File 2. prophage_coordinates.csv

File 3. reads.fastq.1.gz
reads. fastq.2.gz

output/phigaro/

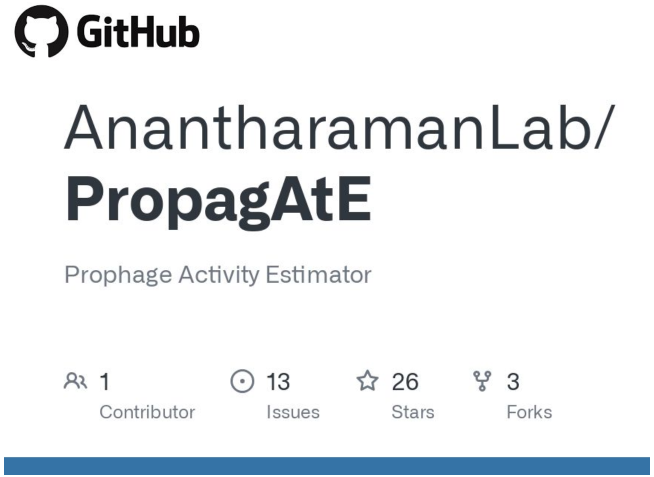


fragment start stop

scaffold ~~id begin end~~ transposable taxonomy

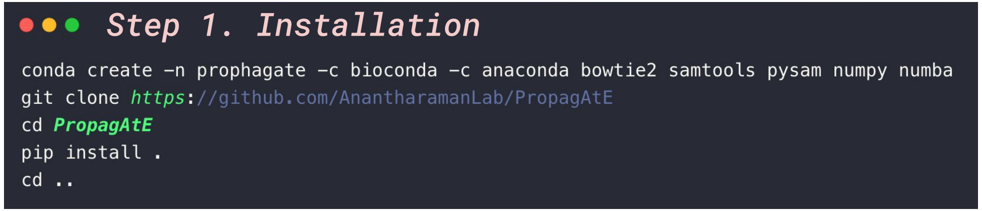
```
AP031427.1  AP031427.1_prophage1    257823  268943  False   Unknown
AP031427.1  AP031427.1_prophage2    342830  348775  False   Unknown
AP031427.1  AP031427.1_prophage3    997202  1015857 False   Siphoviridae
AP031427.1  AP031427.1_prophage4    1655399 1697744 True    Siphoviridae
AP031427.1  AP031427.1_prophage5    1994134 2007394 False   Siphoviridae
```

https://doi.org/10.1128/msystems.00084-22



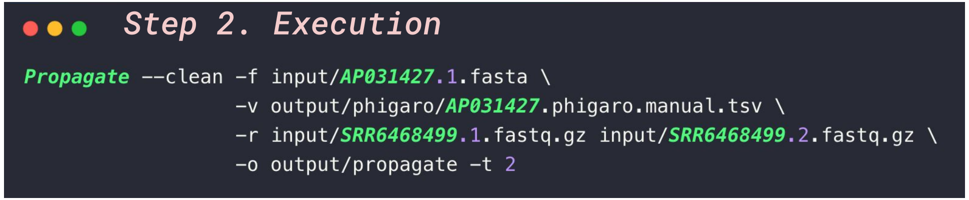
AnantharamanLab/ **PropagAtE**

Prophage Activity Estimator



Step 1. Installation

```
conda create -n prophagate -c bioconda -c anaconda bowtie2 samtools pysam numpy numba
git clone https://github.com/AnantharamanLab/PropagAtE
cd PropagAtE
pip install .
cd ..
```



Step 2. Execution

```
Propagate --clean -f input/AP031427.1.fasta \
                  -v output/phigaro/AP031427.phigaro.manual.tsv \
                  -r input/SRR6468499.1.fastq.gz input/SRR6468499.2.fastq.gz \
                  -o output/propagate -t 2
```

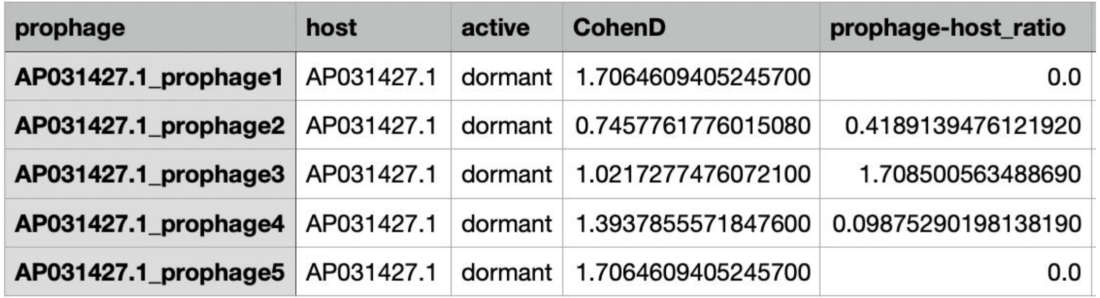
input/

File 1. genome.fasta

File 2. prophage_coordinates.csv

File 3. reads.fastq.1.gz
reads. fastq.2.gz

output/propagate/



prophage	host	active	CohenD	prophage-host_ratio
AP031427.1_prophage1	AP031427.1	dormant	1.7064609405245700	0.0
AP031427.1_prophage2	AP031427.1	dormant	0.7457761776015080	0.4189139476121920
AP031427.1_prophage3	AP031427.1	dormant	1.0217277476072100	1.70850563488690
AP031427.1_prophage4	AP031427.1	dormant	1.3937855571847600	0.0987529019813819O
AP031427.1_prophage5	AP031427.1	dormant	1.7064609405245700	0.0

https://doi.org/10.1128/msystems.00084-22

Thanks for coming!