

Agriculture & Agri-Food Canada Phage Genomics Workshop

Andrew M. Kropinski
Department of Pathobiology
University of Guelph, Canada
Email: Phage.Canada@gmail.com



Acknowledgement

We thank Brian Anderson from DNASTAR Inc. for access to the latest version of Lasergene Suite software



Biography

- ☐ Been working on phages since mid-1960s
- ☐ Held academic and government research positions
- ☐ Currently advise students & faculty at the University of Guelph – Adjunct Professor
- ☐ Past Chair, Bacterial and Archaeal Viruses Subcommittee of ICTV
- ☐ Genome Advisor to NCBI
- ☐ Sequenced: >150 phages
- ☐ Default-setting bioanalyst - Online Analysis Tools (<http://molbiol-tools.ca>)

If you're not a programmer...You're not a Bioinformatician! John H.E. Nash (PHAC)

- ☐ Apologies: I am a Windows and Mac person not a Unix/Linux user



Overview

You can:

1. have a commercial company sequence your phage, and assemble its genome
2. submit this sequence to GenBank and receive an "Accession Number"
3. annotate its genome using an online pipeline which will find that most of the genes specify "hypothetical proteins" OR
4. Devote some time to (a) preparing your genome for annotation and (b) carefully scrutinizing the annotations i.e. follow what you learn in this workshop

Overview 2

Laboratory output:

- ☐ High numbers of bacterial and phage genomes
 - Bioinformatician
 - Computing resources
 - Python programs
- ☐ Low number of bacterial and phage genomes
 - Windows, Mac and internet resources as described in this workshop

Workshop Outline

- ☐ Phage genome assembly – emphasis on Illumina paired-end data
- ☐ Autoannotation coupled with manual proofreading
- ☐ Phage taxonomy

Objectives

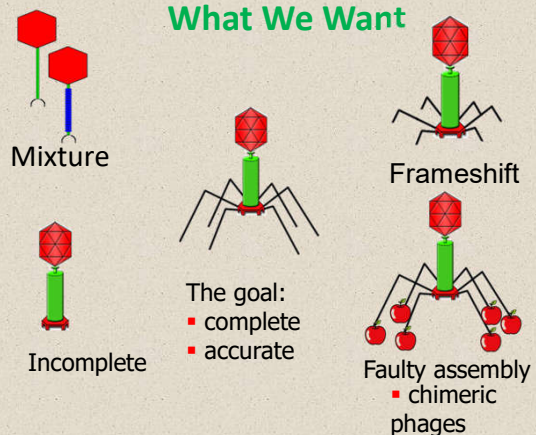
By the end of this workshop participants will

- ☐ have a deeper understanding of the steps involved in sequencing, assembling, and annotating phage genomes
- ☐ understand how phages are classified
- ☐ have an authoritative list of Internet resources and recommended software (commercial and free) for genome analysis.

PART 1

Genome Assembly

What We Want



Outline

Paired-end Illumina sequence data



Assemble
DNASTAR SeqMan NGen18

Primary assembly - contigs



Trim
DNASTAR SeqMan Ultra
Reassemble

Reassembly



Manipulation – RC, Cut/Paste, Splinting
SeqBuilder Pro

Genome for Annotation

Why emphasis on DNASTAR?

- ☐ Developed by geneticist Fred Blattner and computing science student John Schroeder (1984)
- ☐ I have been using their software since 1995
- ☐ Company responds readily to enquiries
- ☐ Software packages are **really** updated annually
- ☐ Available for Mac and PCs
- ☐ Intuitive software
- ☐ Excellent tutorials/videos

Non-commercial Alternatives

- ☐ SPAdes Genome Assembler
 - works with Ion Torrent, PacBio, Oxford Nanopore, and Illumina paired-end
 - Linux, macOS
 - URL: <https://github.com/ablab/spades>
 - Reference: Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics. 2020 Jun;70(1): e102. doi: 10.1002/cpbi.102. PMID: 32559359.
- ☐ **N.B.** I do not recommend Nanopore for phage sequencing

Setup for sequence assembly & analysis

- ☐ Create a directory with the name of the phage under analysis
- ☐ Create four subdirectories:
 - Data
 - Assembly
 - Reassembly
 - Annotation

Genome assembly using SeqMan NGen18

- ☐ Step-by-Step



SeqMan NGen

Welcome! What do you want to do?

New Assembly

Manage Cloud Assemblies

Variant Analysis / Resequencing

RNA-Seq / Transcriptomics

De Novo Assembly and Finishing

Metagenomics

Variant Call Format (VCF) Files

Combine / Reanalyze Existing Assemblies

Move from screen to screen with NEXT

ABI / Sanger

De novo assembly

Genome finishing - refinement

NGS-Based

De novo assembly

Genome finishing - initial error correction

Genome finishing - refinement

Combined reference-guided/de novo assembly

PacBio / Nanopore

De novo assembly (beta)

De novo assembly and polishing (beta)

NGS polishing of draft genome (beta)

Adding Illumina data

Input sequences

Read technology: illumina ☒ Paired-end data

Experiment setup: Single sample

Sequence File: M-Grey_S6_L001_R1_001.fastq.gz M-Grey_S6_L001_R2_001.fastq.gz

Pair Distance: 1000 1000

Input sequences

Read technology: illumina ☒ Paired-end data

Experiment setup: Single sample

Sequence File: M-Grey_S6_L001_R1_001.fastq.gz M-Grey_S6_L001_R2_001.fastq.gz

Pair Distance: 500 500

N.B. Compressed fastq data

Changing number of reads

Preassembly options

Read Filtering

☒ Maximum total reads: 10,000,000

☐ Remove reads smaller than: 15

Preassembly options

Read Filtering

☒ Maximum total reads: 200,000

☐ Remove reads smaller than: 15

Assembly options and names

Assembly options

Coverage Calculation

☐ Unknown genome length

☒ Estimated genome length 150,000 bp

Assembly options

Coverage Calculation

☐ Unknown genome length

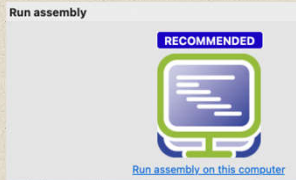
☒ Estimated genome length 150,000 bp

Project name: MGREY2_ASSEMBLY

Project folder: /Users/andrewmkropinski/Desktop/HANY ANANY PHAGES/MGREY2/ASSEMBLY

Assembly output: MGREY2_ASSEMBLY.sqd
MGREY2_ASSEMBLY_contigs.fas
MGREY2_ASSEMBLY.script

Run




Finished

Assembly Log:
 Layout pass 1 - Assemble repeat sequences into repeat contigs
 Assembled: 17825 Unassembled: 20684 Contigs: 30796
 Begin Section: Realign Contigs
 Realigning Contigs
 Realigning contig 30721 of 30796
 Realigning Complete
 Done with alignment
 removeSmallContigs
 MinLength: 100
 saveReport
 file: "I:\Users\andrewm\ropinski\Desktop\HANY ANANY PHAGES\MGREY2\ASSEMBLY\MGREY2_ASSEMBLY\MGREY2_ASSEMBLY.txt"
 saveProject
 file: "I:\Users\andrewm\ropinski\Desktop\HANY ANANY PHAGES\MGREY2\ASSEMBLY\MGREY2_ASSEMBLY\MGREY2_ASSEMBLY_configs.fasta"
 format: Fasta
 Begin Section: Saving Project
 Saving fasta assembly "MGREY2_ASSEMBLY_configs.fasta"
 Saved 14 sequences as project
 file: "I:\Users\andrewm\ropinski\Desktop\HANY ANANY PHAGES\MGREY2\ASSEMBLY\MGREY2_ASSEMBLY\MGREY2_ASSEMBLY.sqf"
 format: SeqMan
 Begin Section: Saving Project
 Saving SeqMan assembly "MGREY2_ASSEMBLY.sqf"
 Saved 47411 sequences and 20 contigs
 Saving SeqMan assembly "MGREY2_ASSEMBLY.sqf" complete.
 writeUnassembledSeqs
 file: "I:\Users\andrewm\ropinski\Desktop\HANY ANANY PHAGES\MGREY2\ASSEMBLY\MGREY2_ASSEMBLY\MGREY2_unasm.fasta"
 saveUnassembledSeqs
 closeProject
 Script Complete
 Assembly finished successfully.



Useful data

Choose one or more analysis options:


Open assembly 


Open Project Report

☒ Project Report ☐ Warnings


SeqMan NGen Assembly Report
 SeqMan NGen 17.6.2 build 9
 Assembly Time: 0:11:25

Assembly Totals

Contigs: 20 

Contigs > 2K: 11 

Contigs To Reach Genome Length 150000: 21

Contigs removed due to small size: 30776 


Assembled Sequences: 47411

Unassembled Sequences: 162589

Sequences not assembled due to complete trimming: 441

Sequences removed due to small contig size: 131464

All Sequences: 200000

Contig N50: 24 Kbases 

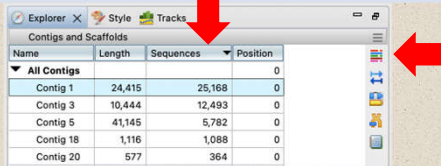
Average Coverage: 87

Examining & editing data in SeqMan Ultra

MGREY2_ASSEMBLY.sqd

/Users/andrewmkropinski/Desktop/HANY ANANY PHAGES/MGREY2/ASSEMBLY/MGREY2_ASSEMBLY

Contig N50: 24.0 kb
Largest contig size: 41,145 bp
Number of contigs: 20
Reads assembled: 47,411

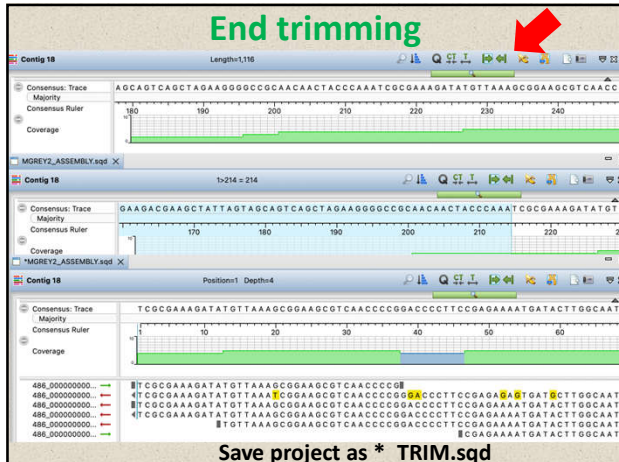


Name	Length	Sequences	Position
All Contigs			0
Contig 1	24,415	25,168	0
Contig 3	10,444	12,493	0
Contig 5	41,145	5,782	0
Contig 18	1,116	1,088	0
Contig 20	577	364	0

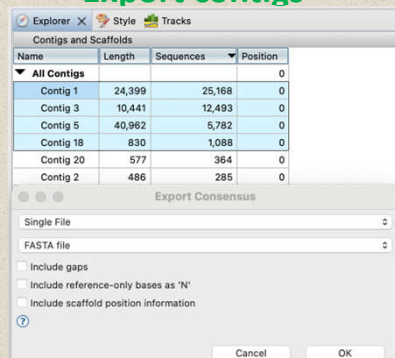
Click here to order by number of sequences

Click here to see a specific contig

End trimming



Export contigs



Explorer X Style Tracks

Contigs and Scaffolds

Name	Length	Sequences	Position
All Contigs			0
Contig 1	24,399	25,168	0
Contig 3	10,441	12,493	0
Contig 5	40,962	5,782	0
Contig 18	830	1,088	0
Contig 20	577	364	0
Contig 2	486	285	0

Export Consensus

Single File ☐

FASTA file ☐

☐ Include gaps

☐ Include reference-only bases as 'N'

☐ Include scaffold position information

Cancel OK

☐ Select and export consensus of desired contigs as a single file in fasta format (*.fas)

Reassembly in SeqMan Ultra

Project Parameters

Assembly

Assembly parameters control the classic assembler method.

- ☒ Consensus
- ☒ Strategy Viewing and Coverage
- ☒ Pair Specifier
- ☒ Assembly
- ☒ End Trimming
- ☒ Variant Discovery

Match Size: 12 (12)

Minimum Match Percentage: 80 (80)

Minimum Sequence Length: 100 (100)

Maximum Added Gaps per kb in Contig: 70 (70)

Maximum Added Gaps per kb in Sequence: 70 (70)

Maximum Register Shift Difference: 70 (70)

Last Group Considered: 2 (2)

Gap Penalty: 0.00 (0.0)

Gap Length Penalty: 0.70 (0.7)

☐ Allow adding single sequence contigs

Only situation where you will want to change the default assembly parameters

Click to select

Reassembly 2

Project Report

Time to Preassemble: 0:0:0

Classic assembly Parameters:

- Match Size: 12
- Minimum Match Percentage: 80
- Minimum Sequence Length: 100
- Maximum Added Gaps per kb in Contig: 70
- Maximum Added Gaps per kb in Sequence: 70
- Maximum Register Shift Difference: 70
- Lastgroup Considered: 2
- Gap Penalty: 0.00
- Gap Length Penalty: 0.70

Entering 4 sequences on 1/21/25.

CREATING NEW Contig 1: from MGRY2_ASSEMBLY_Contig_1(1:24317)

NOT ENTERING in contig 1: MGRY2_ASSEMBLY_Contig_5(1:40950) due to percent match (31)

CREATING NEW Contig 2: from MGRY2_ASSEMBLY_Contig_3(1:10401)

NOT ENTERING in contig 2: MGRY2_ASSEMBLY_Contig_18(1:40950) due to percent match (56)

NOT ENTERING in contig 2: MGRY2_ASSEMBLY_Contig_18(1:40950) due to percent match (56)

ENTERING MGRY2_ASSEMBLY_Contig_18(1:40950) in Contig 1: percent match 100

Time to Assemble: 0:0:1

Examining alignments

Magnification bar

Alignment view

Error check

Consensus: Trace

Majority

Consensus: Ruler

Coverage

Position: 24,151 Depth: 2

Consensus: Trace

Majority

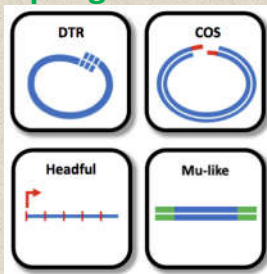
Consensus: Ruler

Coverage

Position: 24,151 Depth: 2

Genome orientation - termini

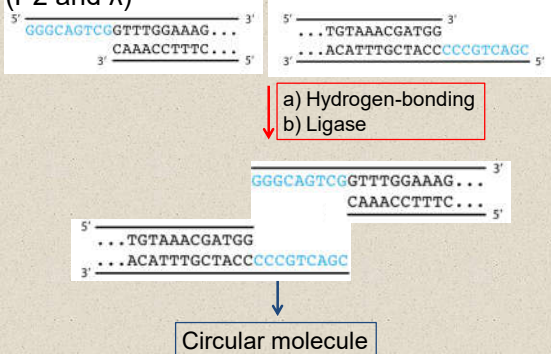
Bacteriophage Genome Termini



- ☐ Phage termini are problematic
- ☐ Bottom line – if your phage is related to an existing virus its genome should be colinear
- ☐ If a choice exists use RefSeq/ICTV homologs
- ☐ **Next six slides are largely for information only**

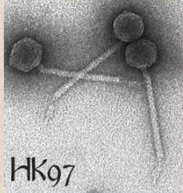
Termini of *Caudoviricetes*

A. Cohesive (sticky) ends – 5'-extensions (P2 and λ)



Termini of Caudoviricetes 2

B. Cohesive (sticky) ends – 3'-extensions
(HK97, D3, many mycobacteriophages)

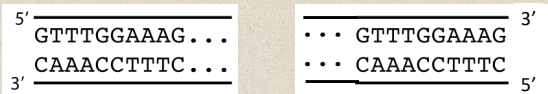


Site-specific packaging

(<http://www.pitt.edu/~duda/HK97.html>)

Termini of Caudoviricetes 3

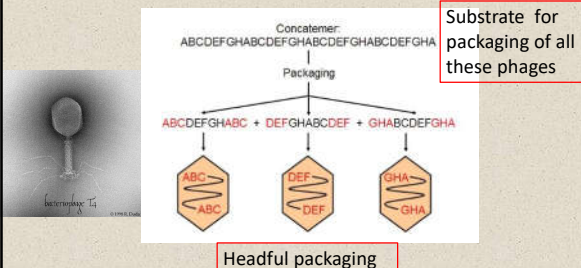
C. Terminal redundancy (TR; direct repeats (DR)) –
Autographivirinae (T7, SP6, ϕ KMV), T5 phage,
Listeria phage A511



- *Escherichia* phage T7 – 160 bp
- *Listeria* phage A511 – 3,125 bp
- *Escherichia* phage T5 – 10,219 bp
- *Bacillus* phage SPO1 – 13,185 bp

Termini of Caudoviricetes 4

D. Terminal redundancy (TR) with Circular Permutation (CP) – *Escherichia* phages T4 and P1

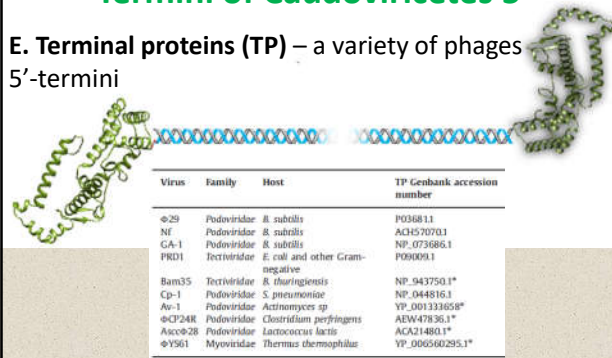


Headful packaging

N.B. No member of the class Caudoviricetes possess a circular genome

Termini of Caudoviricetes 5

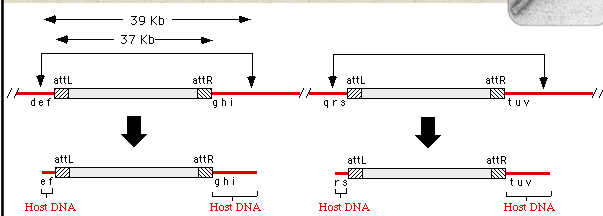
E. Terminal proteins (TP) – a variety of phages 5'-termini



(<http://www.sciencedirect.com/science/article/pii/S0042682214003742>)

Termini of Caudoviricetes 6

F. Host DNA – coliphages Mu & D108



(<http://www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/transposons/Mu.html>)
http://2012.igem.org/Team:Marburg_SYNMIKRO/Project

Some general rules on genome termini

- ☐ T4-like phages begin with rIIA gene on complementary strand
- ☐ Many other phages begin with TerS/L

Termini of Caudoviricetes

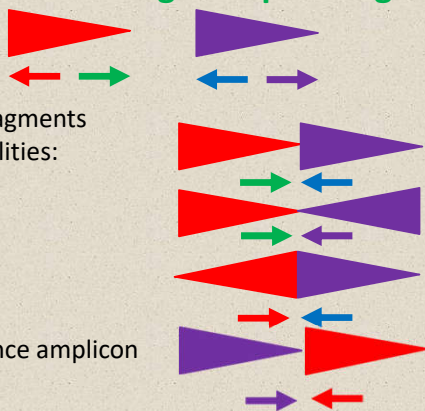
- ☐ PhageTerm - Garneau JR et al. 2017. Sci Rep. **7(1)**:8292. doi: 10.1038/s41598-017-07910-5. PMID: 28811656.
- ☐ accessible via Galaxy Pasteur (<https://galaxy.pasteur.fr/>)
- ☐ video "How to run PhageTerm tool in Galaxy" <https://www.youtube.com/watch?v=9y2gfUSLkgg>
- ☐ for terminal repeats you can use the magnification bar in DNASTAR

What to do about gaps

- ☐ Occasionally assembly result in two or more contigs which will not collapse into one.
- ☐ Gap closure techniques:
 - Primer walking and Sanger sequencing – requires specialize sequencing abilities
 - PCR and Sanger sequencing
 - Splinting and reference-guided assembly

PCR and Sanger sequencing

- ☐ Two fragments
- ☐ Possibilities:
- ☐ Sequence amplicon



Splinting and reference-guided assembly

sequenceNNNNNsequence
(pseudogenome)
BLASTn versus
Caudoviricetes (taxid:
2731619); exclude -
Uncultured/envirom

Teseptimavirus T7 genome assembly, chromosome: 1
Sequence ID: OZ035731.1 Length: 39778 Number of Matches: 2

BLASTn results & splint

Query	421	AGACACCTAAAGTTGGTGGTATCTTTAAGAAGCCTAAGAACAAGGCACAGCGAGAAGGCC	480
Sbjct	18476	AGACACCTAAAGTTGGTGGTATCTTTAAGAAGCCTAAGAACAAGGCACAGCGAGAAGGCC	18535
Query	481	GTGAGCCTTGCGAACTTGAT	500
Sbjct	18536	GTGAGCCTTGCGAACTTGAT	18555

Query	511	GGTGGGTCCCACCAAGTACACCGATAAGGGTGCTCCTGTGGTGGACGATGAGGTACTCG	570
Sbjct	18656	GGTGGGTCCCACCAAGTACACCGATAAGGGTGCTCCTGTGGTGGACGATGAGGTACTCG	18715

- ☐ gap in homologous genome of 99 bp (18556-18655)
- ☐ splint – homologous regions (200 bp) plus missing sequence
- ☐ 18356 - 18855

Splint

Teseptimavirus T7 genome assembly, chromosome: 1
GenBank: OZ035731.1
FASTA Graphics

Teseptimavirus T7 genome assembly, chromosome: 1
GenBank: OZ035731.1
FASTA Graphics

Change region shown
☐ Whole sequence
☒ Selected region
from: begin to: end
Update View

Change region shown
☐ Whole sequence
☒ Selected region
from: 18356 to: 18855
Update View

Teseptimavirus T7 genome assembly, chromosome: 1
GenBank: OZ035731.1
FASTA Graphics

LOCUS OZ035731 500 bp DNA linear PH0 25-APR-2024
DEFINITION Teseptimavirus T7 genome assembly, chromosome: 1.
ACCESSION OZ035731 REGION: 18356..18855

What next

- ☐ use splint to generate a single contig
- ☐ use this contig as the template in a reference-guided assembly with your Illumina sequence data.
- ☐ you will be able to export the consensus for further manipulations and analyses.

Nonaligned contigs

- ☐ what are they and should I worry?
 - Unaligned fragments of your phage
 - Host DNA
 - Prophage DNA
 - Second phage

Is my phage related to anything?

The screenshot shows the NCBI BLAST search page. Red arrows point to the following elements:

- The **blastn** tab at the top.
- The **Enter Query Sequence** text box.
- The **Choose Search Set** section, specifically the **Database** dropdown (set to 'Core nucleotide database (core_nt)') and the **Organism** dropdown (set to 'Caudoviricetes (taxid:2731619)').
- The **Program Selection** section, specifically the **Optimize for** radio buttons (set to 'Somewhat similar sequences (blastn)').
- The **BLAST** button at the bottom left.
- The checkbox **Show results in a new window** at the bottom.

Is my phage related to anything 2?

Pseudomonas phage vB_Pae_LESphi2, complete genome

GenBank: OQ594955.1

[FASTA](#) [Graphics](#)

[Go to:](#) ☐

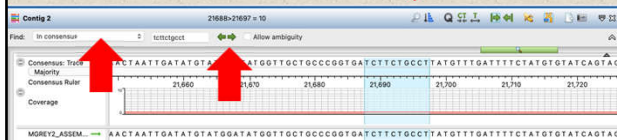
LOCUS OQ594955 42123 bp DNA linear PHG 08-MAR-2024
 DEFINITION Pseudomonas phage vB_Pae_LESphi2, complete genome.
 ACCESSION OQ594955
 VERSION OQ594955.1
 KEYWORDS .
 SOURCE Pseudomonas phage vB_Pae_LESphi2
 ORGANISM [Pseudomonas phage vB_Pae_LESphi2](#)
 Viruses; Duplodnaviria; Heunggongvirae; Uroviricota;
 Caudoviricetes.

Realigning your genome 2

- ☐ May require RC
- ☐ May require cutting and reassembly using SeqMan Ultra
- ☐ Requires SeqBuilder Pro



ORIGIN
 1 tcttctgcct tatgtttgat ttctatgtg tatcagtagc taagactga tcgcttccac



- ☐ Results in your phage genome being colinear with this phage

Is phage ready for annotation?

- ☐ Questions:
 1. Is it full length?
 2. Is it error free?
- ☐ Quick and dirty approaches:
 1. BLASTn
 2. BLASTx

BLASTn – complete example?

LOCUS PP039555 39435 bp DNA linear PHG 18-SEP-2024
 DEFINITION Enterobacter phage vB_Ecl_MII_004, complete genome.
 ACCESSION PP039555
 VERSION PP039555.1
 KEYWORDS
 SOURCE Enterobacter phage vB_Ecl_MII_004
 ORGANISM Enterobacter phage vB_Ecl_MII_004
 Viruses.
 REFERENCE 1 (bases 1 to 39435)

- ☐ BLASTn reveals it is complete and related to *Escherichia* phage Peacock (MK903279; 39233 bp)
- ☐ Caudoviricetes; Autographiviridae; Studiervirinae; Kayfunavirus

BLASTn – incomplete?

LOCUS PP935704 15914 bp DNA linear PHG 03-AUG-2024
 DEFINITION Salmonella phage vB_SE126_2P, complete genome.
 ACCESSION PP935704
 VERSION PP935704.1
 KEYWORDS
 SOURCE Salmonella phage vB_SE126_2P
 ORGANISM Salmonella phage vB_SE126_2P
 Viruses.
 REFERENCE 1 (bases 1 to 15914)

- ☐ BLASTn reveals it is incomplete

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Salmonella phage 118SE, complete genome	Salmonella phage 118SE	21803	22575	100%	0.0	91.80%	41868	NC_028698
<input checked="" type="checkbox"/>	Salmonella phage 12SE, complete genome	Salmonella phage 12SE	21599	22572	100%	0.0	91.89%	41865	KJ051146.1
<input checked="" type="checkbox"/>	Salmonella phage 13SE, complete genome	Salmonella phage 13SE	21594	22568	100%	0.0	91.89%	41867	KJ051147.1
<input checked="" type="checkbox"/>	Salmonella phage 15E4S, complete genome	Salmonella phage 15E4S	21581	22554	100%	0.0	91.87%	41768	KT881477.1
<input checked="" type="checkbox"/>	Salmonella phage 15E1C, complete genome	Salmonella phage 15E1C	21547	22520	100%	0.0	91.81%	41720	KT962832.1
<input checked="" type="checkbox"/>	Salmonella phage vB_SenS_PVP-SE2, complete genome	Salmonella phage vB_SenS_PVP-SE2	18206	22565	100%	0.0	91.92%	42425	NC_073186

BLASTx - frameshifts

LOCUS PP039555 39435 bp DNA linear PHG 18-SEP-2024
 DEFINITION Enterobacter phage vB_Ecl_MII_004, complete genome.
 ACCESSION PP039555
 VERSION PP039555.1

- ☐ BLASTn reveals it is complete and related to *Escherichia* phage Peacock (MK903279; 39233 bp)
- ☐ BLASTx versus *Escherichia* phage Peacock (taxid:2591100) – **huge number of frameshifts**

Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200 ?

Distribution of the top 119 Blast Hits on 47 subject sequences



What are the problems

- ☐ scientists with little knowledge and experience working with phages
- ☐ individuals not taking advantage of free expertise in the International Committee on Taxonomy of Viruses (ICTV)



- ☐ Nanopore versus Illumina sequencing technology

```

COMMENT  ##Assembly-Data-START##
Assembly Method      :: Canu v. 1.7.1; Flye v. 2.9; Racon v.
                  1.4.13
Coverage             :: 24.22x
Sequencing Technology :: Oxford Nanopore Technology
##Assembly-Data-END##
FEATURES
Location/Qualifiers
  
```

End of Part 1



Questions?
