Agriculture & Agri-Food Canada Phage Genomics Workshop

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

Agriculture and Agri-Food Canada
Agriculture et Agroalimentaire Canada

UNIVERSITY of GUELPH

Acknowledgement

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

DNASTAR

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

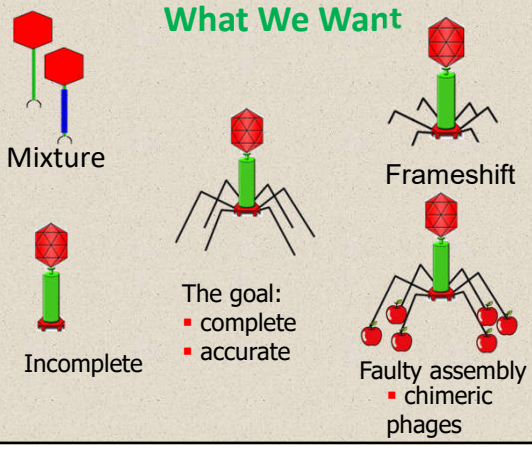
By the end of this workshop participants will

- ❑ 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

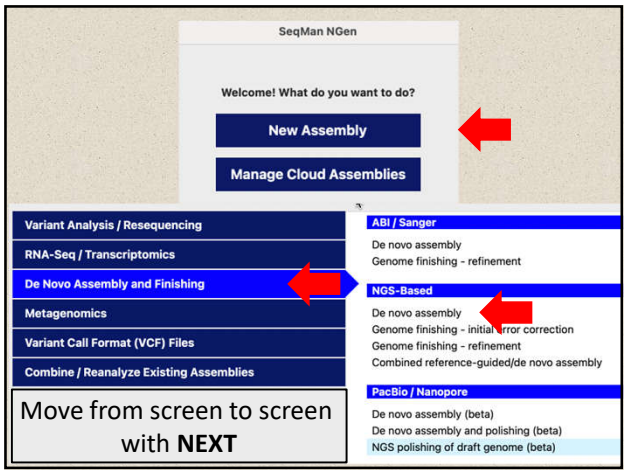
- ❑ SPAdes Genome Assembler
  - ➢ works with Ion Torrent, PacBio, Oxford Nanopore, and Illumina paired-end
  - ➢ Linux, macOS
  - ➢ URL: https://github.com/ablab/spades
  - ➢ Reference: Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics. 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

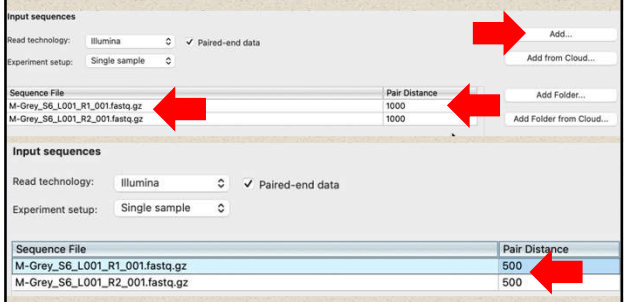
- ❑ Create a directory with the name of the phage under analysis
- ❑ Create four subdirectories:
  - ➢Data
  - ➢Assembly
  - ➢Reassembly
  - ➢Annotation

Genome assembly using SeqMan NGen18

- ❑ Step-by-Step

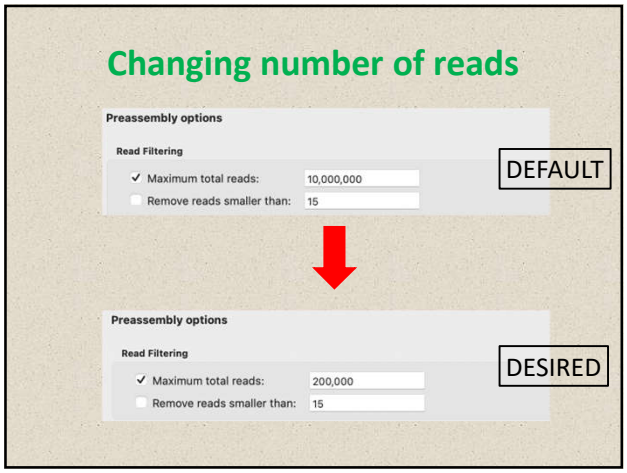


Adding Illumina data

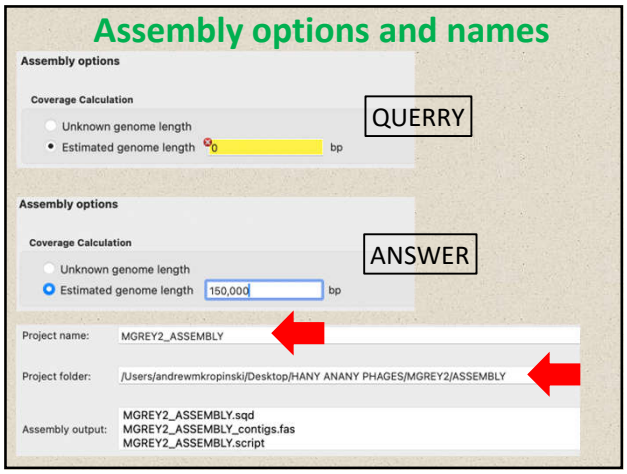


N.B. Compressed fastq data

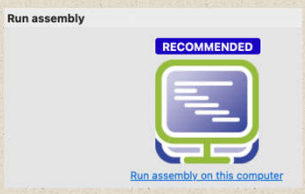
Changing number of reads



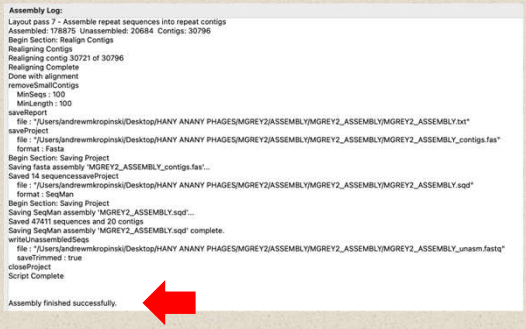
Assembly options and names



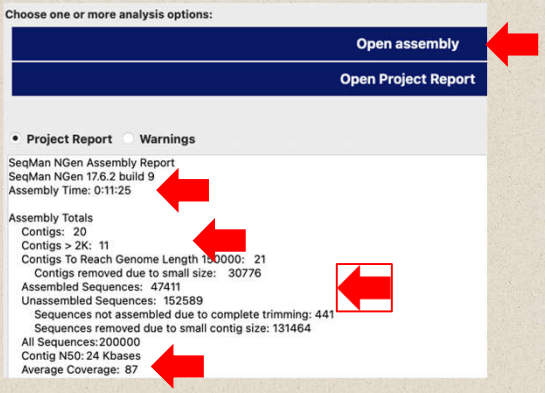
Run



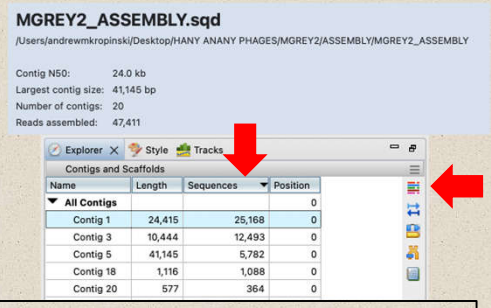
Finished



Useful data



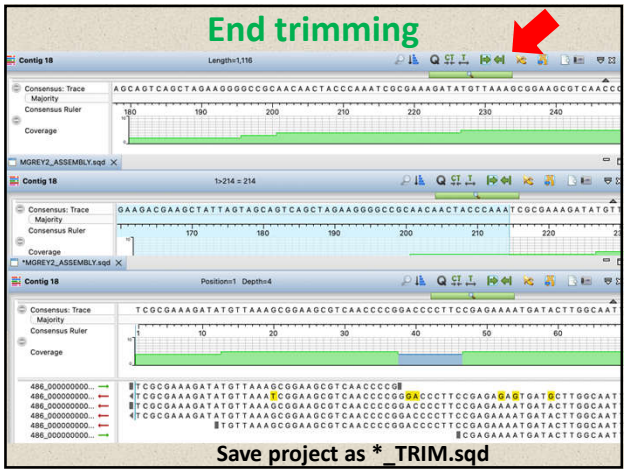
Examining & editing data in SeqMan Ultra



Click here to order by number of sequences

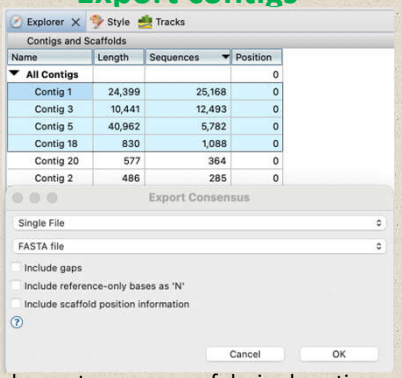
Click here to see a specific contig

End trimming



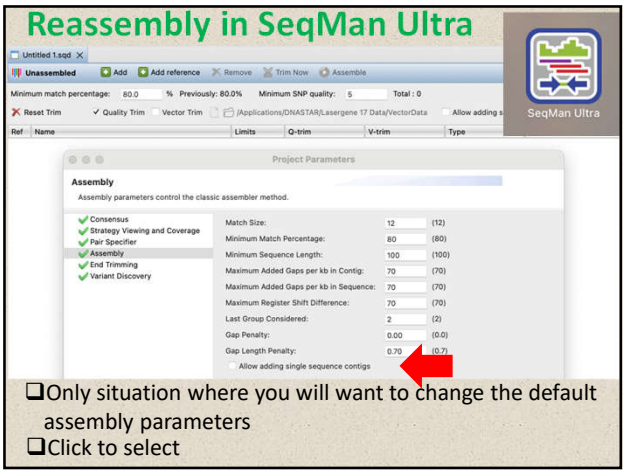
Save project as *_TRIM.sqd

Export contigs



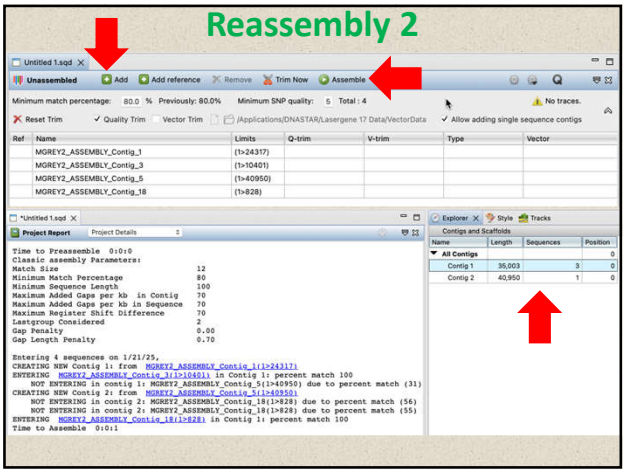
❑Select and export consensus of desired contigs as a single file in fasta format (*.fas)

Reassembly in SeqMan Ultra

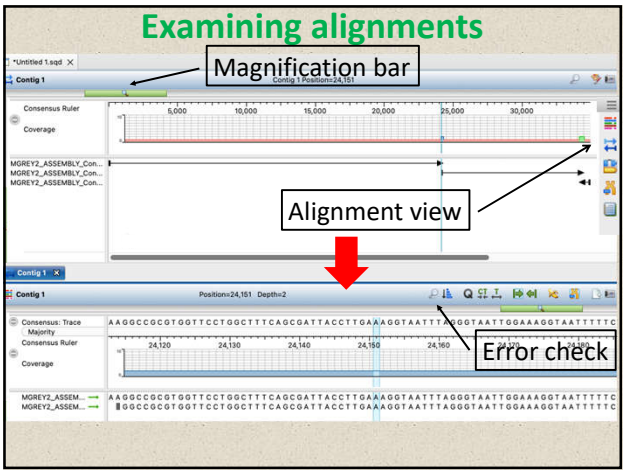


- Only situation where you will want to change the default assembly parameters
- Click to select

Reassembly 2

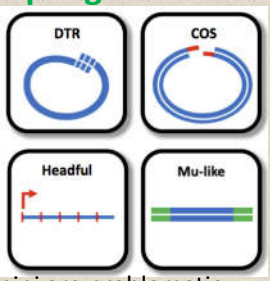


Examining alignments



Genome orientation - termini

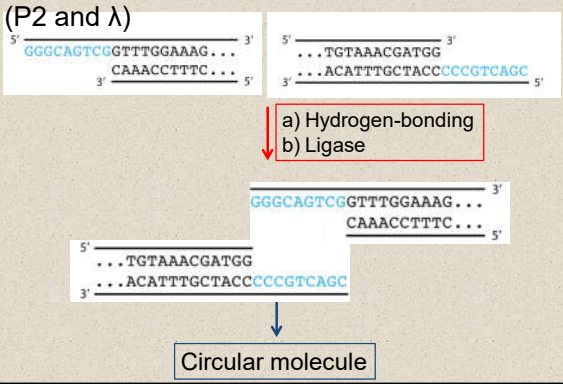
Bacteriophage Genome Termini



- ❑ 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**

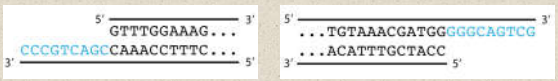
Termini of *Caudoviricetes*

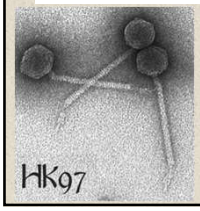
A. Cohesive (sticky) ends – 5'-extensions (P2 and λ)



Termini of Caudoviricetes 2

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



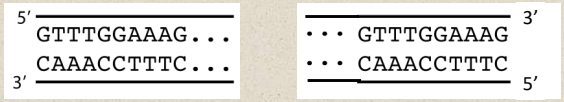


Site-specific packaging

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

Termini of Caudoviricetes 3

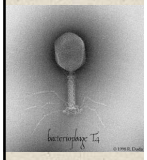
C. Terminal redundancy (TR; direct repeats (DR)) – *Autographivirinae* (T7, SP6, ϕKMV), T5 phage, *Listeria* phage A511

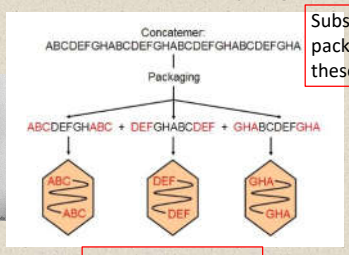


- *Escherichia* phage T7 – 160 bp
- *Listeria* phage A511 – 3,125 bp
- *Escherichia* phage T5 – 10,219 bp
- *Bacillus* phage SPO1 – 13,185 bp

Termini of Caudoviricetes 4

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





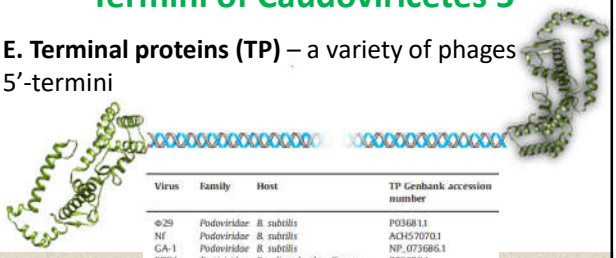
Substrate for packaging of all these phages

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

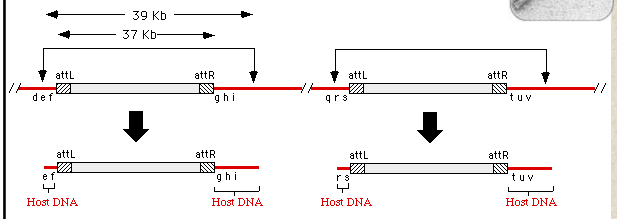


Virus	Family	Host	TP Genbank accession number
Φ29	*Podoviridae*	*B. subtilis*	P03681.1
Nf	*Podoviridae*	*B. subtilis*	ACH57070.1
GA-1	*Podoviridae*	*B. subtilis*	NP_073686.1
PRD1	*Tectiviridae*	*E. coli* and other Gram-negative	P09009.1
Bam35	*Tectiviridae*	*B. thuringiensis*	NP_943750.1*
Cp-1	*Podoviridae*	*S. pneumoniae*	NP_044816.1
Av-1	*Podoviridae*	*Actinomyces sp*	YP_001133658*
ΦCP24R	*Podoviridae*	*Clostridium perfringens*	AEW47836.1*
Asco-28	*Podoviridae*	*Lactococcus lactis*	ACA21480.1*
ΦYS61	Myoviridae	*Thermus thermophilus*	YP_006560295.1*

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

Termini of Caudoviricetes (

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

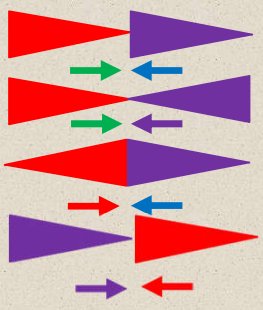
- PhageTerm - Garneau JR et al. 2017. Sci Rep. **7(1):**8292. doi: 10.1038/s41598-017-07910-5. PMID: 28811656.
- accessible via Galaxy Pasteur (https://galaxy.pasteur.fr/)
- video "How to run PhageTerm tool in Galaxy" https://www.youtube.com/watch?v=9y2gfUSLkgg
- for terminal repeats you can use the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

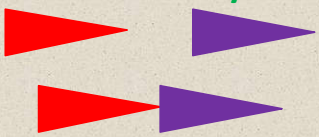
PCR and Sanger sequencing

- Two fragments
- Possibilities:



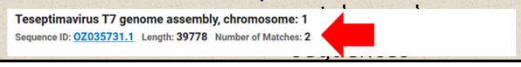
- Sequence amplicon

Splinting and reference-guided assembly

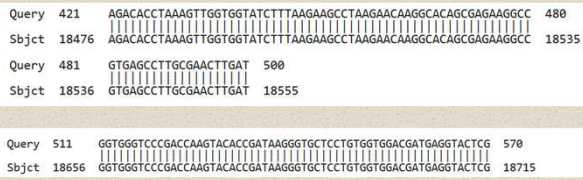


sequenceNNNNNsequence
(pseudogenome)

BLASTn versus Caudoviricetes (taxid: 2731619); exclude - Uncultured/environm

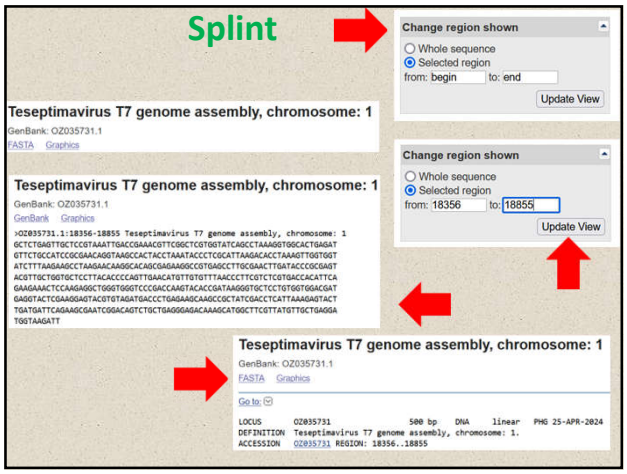


BLASTn results & splint



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

Splint



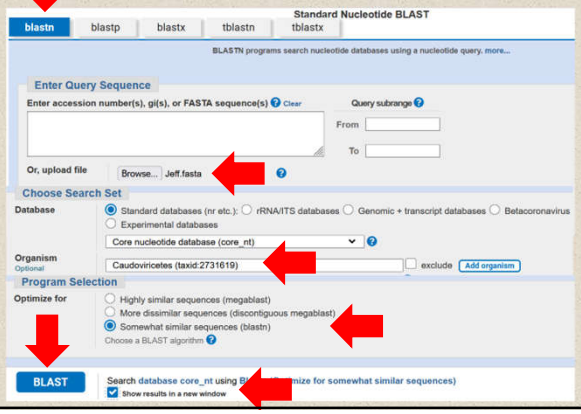
What next

- ❑ 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?



Is my phage related to anything 2?

Pseudomonas phage vB_Pae_LESphi2, complete genome

GenBank: OQ594955.1

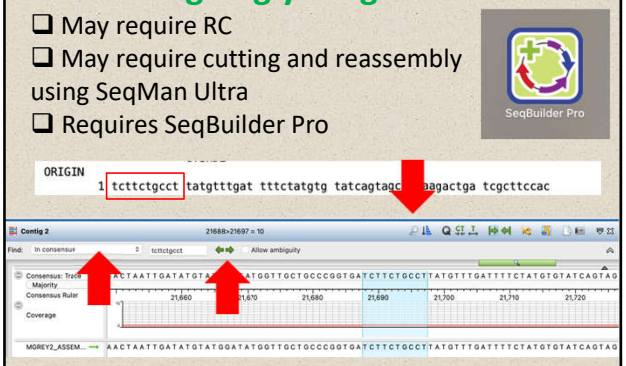
FASTA Graphics

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



- ❑ 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       PQ039555               39435 bp    DNA     linear   PHG 18-SEP-2024
DEFINITION  Enterobacter phage vB_Ecl_MII_004, complete genome.
ACCESSION   PQ039555
VERSION     PQ039555.1
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)
```

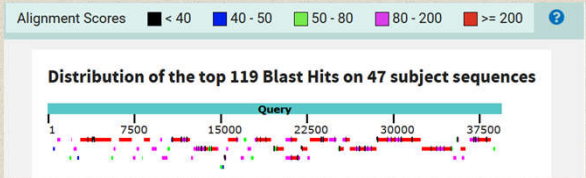
- ❑ BLASTn reveals it is incomplete

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Salmonella phage f18SE, complete genome	Salmonella phage f18SE	21603	22575	100%	0.0	91.90%	41868	NC_028698.
Salmonella phage f2SE, complete genome	Salmonella phage f2SE	21599	22572	100%	0.0	91.89%	41865	KU951146.1
Salmonella phage f3SE, complete genome	Salmonella phage f3SE	21594	22568	100%	0.0	91.89%	41867	KU951147.1
Salmonella phage fSE4S, complete genome	Salmonella phage fSE4S	21581	22554	100%	0.0	91.87%	41768	KT881477.1
Salmonella phage fSE1C, complete genome	Salmonella phage fSE1C	21547	22520	100%	0.0	91.81%	41720	KT962832.1
Salmonella phage vB_SenS_PVP-SE2, complete genome	Salmonella phage vB_SenS_PVP-SE2	18266	22585	100%	0.0	91.92%	42425	NC_073186.

BLASTx - frameshifts

```
LOCUS       PQ039555               39435 bp    DNA     linear   PHG 18-SEP-2024
DEFINITION  Enterobacter phage vB_Ecl_MII_004, complete genome.
ACCESSION   PQ039555
VERSION     PQ039555.1
```

- ❑ 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**



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?