

Agriculture & Agri-Food Canada Phage Genomics Workshop

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






PART 2

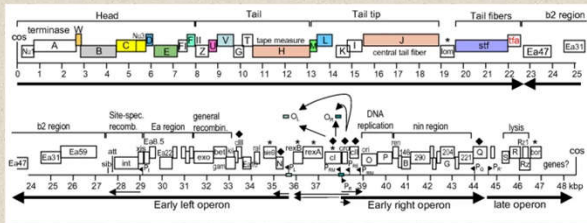
Genome annotation

APOLOGY: NO AUTOANNOTATION SOFTWARE PACKAGE IS PERFECT. YOU WILL NEED TO DO A LOT OF PROOFREADING TO ACHIEVE A HIGH-QUALITY ANNOTATED VIRAL GENOME

Outline

- ☐ Online primary genome annotation with DFAST, PHAROKKA or RAST – **PART 2A** 
- ☐ Massaging *.gbk file
- ☐ Identification of missing coding sequences using UGENE or Geneious Prime – **PART 2B** 
- ☐ GenBank file types & their interconversion – **PART 2C**
- ☐ Examination of proteins for conserved motifs – **PART 2D** 
- ☐ Updated GenBank flatfile (*.gbk)

Annotation outcome



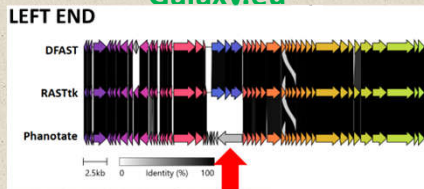
- ☐ We want to identify the location and function of all the genes on our phage genome:
- ☐ Two main classes encode proteins and tRNAs

Automated Annotation – PART 2A

- ☐ Proksee - <https://proksee.ca/> uses Prokka
- ☐ RAST - Rapid Annotation using Subsystem Technology (<https://rast.nmpdr.org/>)
- ☐ DFAST - DDBJ Fast Annotation and Submission Tool. (<https://dfast.nig.ac.jp/>)
- ☐ BV-BRC - Bacterial and Viral Bioinformatics Resource Center (<https://www.bv-brc.org/app/Annotation>)
- ☐ GenSAS v6.0 - <https://www.gensas.org/>
- ☐ MicroScope Microbial Genome Annotation & Analysis Platform - <https://mage.genoscope.cns.fr/microscope/home/index.php>
- ☐ Bakta Web - <https://bakta.computational.bio/>

All except DFAST & Bakta require free online registration

Use Pharokka with caution at Galaxy.eu



First pass annotation using DFAST

DFAST

- ☐ DDBJ Fast Annotation and Submission Tool
- ☐ associated with DNA Data Bank of Japan
- ☐ Tanizawa Y et al. 2018. Bioinformatics. **34(6)**: 1037-1039. doi: 10.1093/bioinformatics/btx713. PMID: 29106469
- ☐ <https://dfast.ddbj.nig.ac.jp/>
- ☐ click on "Start your project"

Input

DFAST Prokaryotic genome annotation pipeline

Query File (Fasta format, up to 15Mbyte)

No file selected.

Name/Title for the Job (optional)

General Settings

Organism (optional)

Strain (optional)

Locus tag prefix [?](#) LOCUS

Minimum sequence length 200

Advanced Options

General Settings

Organism Pseudomonas phage Jeff

Strain (optional)

Locus tag prefix [?](#) JEFF

Minimum sequence length 75

☐ Name of your phage - JEFF

☐ Institute + # - GRDC57

AMR/Virulence Factors

Antimicrobial Resistance Genes and Virulence Factors

[Beta] Annotation of Antimicrobial Resistance Genes and Virulence Factors using the CARD/VFDB reference databases and identification of plasmid-derived contigs using PlasmidFinder.

☒ Enable annotation of AMR/VFG

Results

☐ 17 seconds latter:

Genome Statistics

Total Length (bp)	66,776
No. of Sequences	1
GC Content (%)	55.6%
N50	66,776
Gap Ratio (%)	0.0%
No. of CDSs	91
No. of rRNA	0
No. of tRNA	0
No. of CRISPRS	0
Coding Ratio (%)	91.6%

Download Files

Genbank Flat File :	annotation.gbk
GFF3-formatted File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Pseudogene Summary :	pseudogene_summary.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip



What do they all mean?

- *.faa Protein FASTA file of the translated CDS sequences.
- *.ffn Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
- *.fna Nucleotide FASTA file.
- *.gbk This is a standard GenBank flatfile
- *.gff This is the master annotation in GFF3 format, containing both sequences and annotations.
- *.tsv Tab-separated file of all features

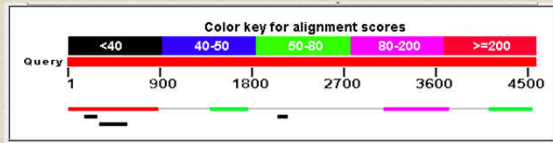
Problems with DFAST

- ☐ it is prokaryote annotation pipeline NOT a phage pipeline, therefore annotations are incomplete and skewed.
- ☐ while you can change the translation code from 11 (default) to 4 you cannot change it to 15 which is used by some gut phages.
- ☐ splicing not recognized:
 - in GenBank record this is reported as CDS join(99941..100807,101606..103957)
 - important with members of the *Herelleviridae*, such as *Staphylococcus* phages; and, *Campylobacter* phages

Introns and Inteins – rare but

- the gene encoding the aerobic ribonucleoside diphosphate reductase (large subunit) from *Campylobacter* phage vB_CcoM-IBB_35 is located on a 4.5 kb region which homologous to a 2.3 kb region from *Klebsiella* phage vB_KleM-RaK2. N.B. proteins have similar mass.

N.B. the thin grey line joining blocks indicates a downstream sequence which is similar



Editing primary annotation (*.gbk) files

- No matter what program you use you will have to edit the results
- Notepad++ is better than Notepad
- Install from: <https://notepad-plus-plus.org/>
- Notes: <https://www.youtube.com/watch?v=wVfbFe57h2o>

Open in Notepad++

- Install from: <https://notepad-plus-plus.org/>

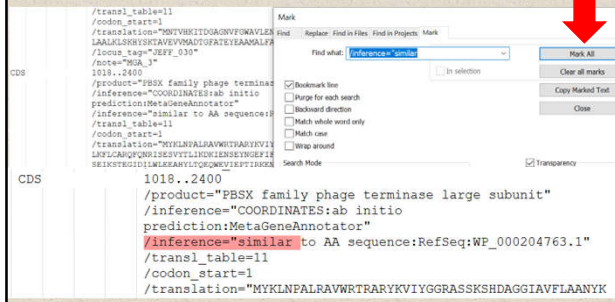
```
CDS      1018..2400
          /product="PBSX family phage terminase large subunit"
          /inference="COORDINATES-ab-initio
          prediction:MetaGeneAnnotator"
          /inference="similar to AA sequence:RefSeq:WP_000204763.1"
          /transl_table=11
          /codon_start=1
          /translation="MYKLNPALRAVWRTRARYKVIYGRASSKSHDAGGIAVFLAANYK
          LKFLCARQFQNRISSEVYTLIKDKIENSEYNGEFIFTKNSIKHKSTGSEFLFYGIARNL
          SEIKSTEGIDILWLEEAHYLTQEQWEVIEPTIRKENSEIWIIFNPNEVDFVYQNFVK
          PPKDSCVKMINNENPFLSETMLKVIHEAYERDREQAEHIYGGIPTGGDKSVINLKFI
          LAADAHKKLWEPAGSKRIGFDVADGDGANATTLMHGNVIMEVDEMDGLELLKSS
          SRVYNLAKMKGASVTYDSIGVGAHVGSKFAELNDASPDFKLIYDPFNAGGAVDKPDDIY
          HKLPHTTIKNDHFSNLIKAGWEEVATRFKTYEAVHGKVPYDELISINSEIHPDK
          LNQLCIELSSPRKLDWNGRFKVESKDKMREKRKIKSPNIADSVIMSAIILIRKPKGFF
          DF"
          /locus_tag="JEFF_040"
          /note="WP_000204763.1 PBSX family phage terminase large
          subunit (Enterobacteriaceae) [pid:54.3%, q_cov:97.4%,
          s_cov:94.9%, Eval:1.3e-133]"
          /note="MSA_4"
          complement(2437..2820)
```

★ Comment

★ Delete

Open in Notepad++

❑ Under: "Search" choose "Mark" and "Mark All"



Mark All/Delete

```
CDS      1018..2400
          /product="PBSX family phage terminase large subunit"
          /inference="COORDINATES:ab initio
          prediction:MetaGeneAnnotator"
          /inference="Similar to AA sequence:RefSeq:WP_000204763.1"
          /transl_table=11
          /codon_start=1
          /translation="MYKLNPALRAVWRTRARYKVIYGRASSKSHDAGGIIVFLAANYK
          LKFLCARQPNRISESVTYLLKDKIENSETWGEFTFTNSINRSTGSEFLYGIARML
          SEIKSTGIDILALEARVLTQKQWVIEPIRKENSETWIIFNNEVDYVQNPVVK
          PFKDSCVMINWNPFLSETMLKVIHEAVERDREQAHIYGGIPIKGGOKSVINLKF
          LAIDAHHKLGWEFAGSKRIGFDVADGDEANATTIMHGNVMEVDWDGLEDELLKSS
          SRVNLAKMKGASVYDSIGVGAHVGSFAELNDASPDKFLIYDFNAGGAVDKPDDIY
          MCLPHTTIKNDHFSNIAQKWEVATRFKTYEAVESGVYFPDELISINSETIHPK
          LNLCLTELSSPRKDLMMGRKRVESKDKMGEKSKIKSPNIADSVMSAILPIPKPKOFF
          DF"
          /locus_tag="JEFF_040"
          /note="WP_000204763.1 PBSX family phage terminase large
          subunit (Enterobacteriaceae) [pid:54.3%, q_cov:97.4%,
          e_cov:194.9%, Eval:1.3e-133]"
          /note="MGA_4"
```

- ❑ Under "Search", choose "Bookmark" and "Remove Bookmarked Lines"
- ❑ It will not remove multiple lines so you will have to manually accomplish this
- ❑ shorten "product" to: terminase, large subunit

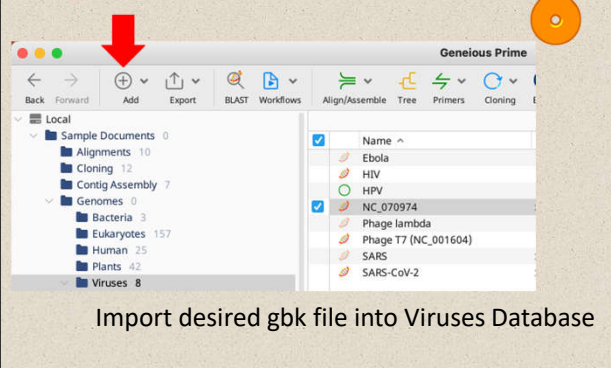
Edit top of *.gbk file

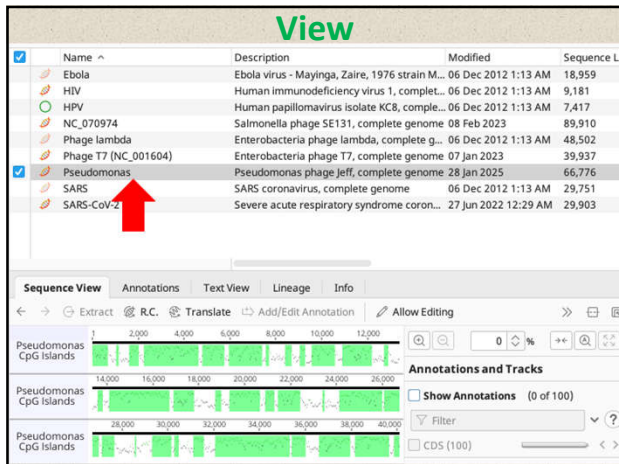
```
LOCUS       sequencel               66776 bp    DNA    linear    BCT 03-FEB-2025
DEFINITION  Pseudomonas phage Jeff unspecified. DNA, contig: sequencel.
ACCESSION   sequencel
VERSION     sequencel.1
KEYWORDS    .
SOURCE      Pseudomonas phage Jeff unspecified.
  ORGANISM  Pseudomonas phage Jeff unspecified.
COMMENT     Annotated by DFAST (https://dfast.nig.ac.jp).
```

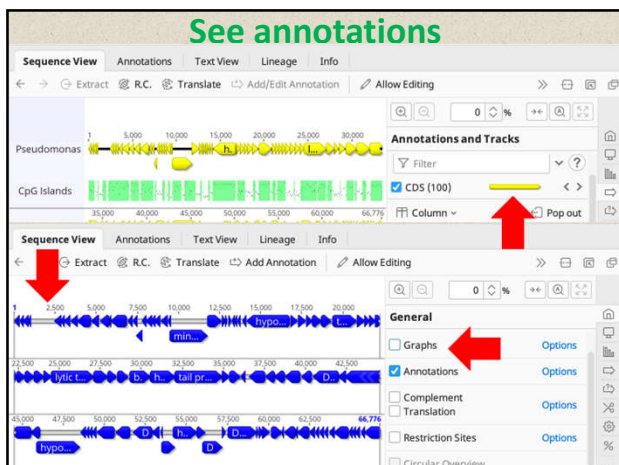
```
LOCUS       Jeff                   66776 bp    DNA    linear    PHG 28-JAN-2025
DEFINITION  Pseudomonas phage Jeff, complete genome.
ACCESSION   .
VERSION     .
KEYWORDS    .
SOURCE      .
  ORGANISM  Pseudomonas phage Jeff
            Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes;
            Pbuonavirus;
FEATURES             Location/Qualifiers
```

This will be discussed in this workshop BUT in many cases the taxonomy can be readily assessed from the BLASTN hits

Visualizing & editing GBK files using Geneious Prime – PART 2B







Find missing ORFs

Minimum size: 228 nt = 75 amino acids plus stop codon

Find missing ORFs 2

☐ BLASTp vs Caudoviricetes (taxid:2731619)

☐ Add new feature to the annotation

OR

U Visualizing & editing GBK files using UGENE – PART 2B

- ❑ Rose R et al. 2019. Bioinformatics **35(11)**: 1963-1965 doi: 10.1093/bioinformatics/bty901. PMID: 30358807.
- ❑ Window, Linux & Mac versions
- ❑ <http://ugene.net/>
- ❑ podcasts: <https://www.youtube.com/@UniproUGENE>

Welcome to UGENE

Open File(s)

UGENE

Select an annotation name:

Annotation: CDS Color: [blue]

Show all annotation names

Click here to change colour

Magnification

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 6 216

Screen for missing genes

- ❑ appear as gaps

Select an annotation name:

Annotation: CDS Color: [blue]

Show all annotation names

Click here to change colour

Magnification

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 6 216

- ❑ Highlight "missing region"

900 [1606 bp] 2505

See sequence data

ACG

CDS (27)

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k

ATTACTCTTCGGCCTTCGTCCGGCGCTCAGCCACTCTTC

Pseudomonas [dna]

CDS (27)

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k

GCCTATTGCGAACAATGGACTGATTAATGACAAAC

GCGATAACGGCTTGTACCTGACTAATTACATGTTTG

Find ORF

Find ORF

Strand

☒ Both

☐ Direct

☐ Complement

Search Settings

☒ Min length, bp: 75

☒ Must terminate within region

☒ Must start with init codon

☒ Allow overlaps

☒ Allow alternative init codons

☒ Include stop codon

☒ Max result: 100

Region Custom region 911 - 2539

Region	Strand	Length
[1018..2400]	Direct	1383
[1039..2400]	Direct	1362
[1327..2400]	Direct	1074

Click here

Find ORF 2

Find ORF 2

CDS (27)

1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k

GCCTATTGCGAACAATGGACTGATTAATGTACAAACTCAA

GCGATAACGGCTTGTACCTGACTAATTACATGTTTGAGTT

Add new annotation

- Under “Actions” in the top bar choose:

➤ “Add” and “New annotation”

Annotation type: bHLH Domain, C_region, CAAT Signal, CDS, Cellular, centromere, Conflict, D-Loop, D_segment, Enhancer, exon, gap, GC-Signal, gene, Glycosylation Site, Homeodomain

Group name: CDS

Annotation name: by type

Description: terminase, large subunit

Location: Simple format (selected), 1018 - 2400, Complement (unchecked), GenBank/EMBL format (selected), 1018..2400

Alternatives

- NCBI's Genome Workbench – Windows, Mac, Ubuntu; discontinued 2024
(<https://ftp.ncbi.nlm.nih.gov/toolbox/gbench/ver-3.9.1/>)

- DNA Master

<https://phagesdb.org/DNAMaster/>



ORF versus CDS

- Open Reading Frames vs Coding Sequences
- DFAST identified 91 CDSs in *Pseudomonas* phage Jeff
- NCBI's Open Reading Frame Finder “ORF finder”
ORFs found: 795 Genetic code: 11 Start codon: 'ATG' and alternative codons



<https://www.ncbi.nlm.nih.gov/orffinder/>

- Basic difference is that CDSs are preceded by ribosome-binding sites

CDS, ORF and Genes - optional

"Most protein sequences are derived from translations of **CoDing Sequence (CDS)** derived from gene predictions. A **CoDing Sequence (CDS)** is a region of DNA or RNA whose sequence determines the sequence of amino acids in a protein. It should not be mixed up with an **Open Reading Frame (ORF)**, which is a series of DNA codons that does not contain any STOP codons. **All CDS are ORFs, but not all ORFs are CDS...**"

(https://www.uniprot.org/help/cds_protein_definition)

ORF vs CDS 2 - optional

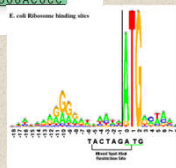
- ☐ an ORF is a sequence that has a length divisible by three and is bounded by stop codons
- ☐ stop codons - TAA, TAG or TGA
- ☐ may not specify a protein

(Sieber P, Platzer M, Schuster S. 2018. The Definition of Open Reading Frame Revisited. Trends in Genetics, 34 (3): 167-170)

CDS - optional

	Shine-Dalgarno sequence	Start codon
<i>E. coli araB</i>	UUUGGAU	GGAGUGAAACG
<i>E. coli lacZ</i>	CAAUUCAGGG	UGGUAAU
<i>E. coli lacZ</i>	UUCACACAGGAAACAGCU	AUGAACCAUGAUU
<i>E. coli thrA</i>	GGUAAACAGGUAACAGGA	AUGCGAGUGUUG
<i>E. coli trpA</i>	AGCACGAGGGGAAUCUG	AUGGAACGCUAC
<i>E. coli trpB</i>	AUAUGAAGGA	AAGGAACA
<i>λ</i> phage cro	AUGUACUAAGGAGGUUGU	AUGGAACACGC
R17 phage A protein	UCCUAGGAGGUUUGACCU	AUGGAGCUUUU
φJ phage A replicase	UAAUAAGGAUGAAU	GC AUGUCUAAAGCA
φX174 phage A protein	AAUCUUGGAGGCUUUUUU	AUGGUUCGUUCU
<i>E. coli</i> RNA polymerase B	AGCGAGCU	GAGGAACCCU

E. coli Ribosome binding sites



N.B. No matter what the "start" codon it is always translated as M (Met, methionine)

File format interconversions - PART 2C

- ☐ Genome2D – conversion
- ☐ Baerends RJ et al. 2004. Genome Biol. **5(5)**: R37. doi: 10.1186/gb-2004-5-5-r37. PMID: 15128451.
- ☐ http://genome2d.molgenrug.nl/g2d_tools/conversions.html
- ☐ **Warning:** is the conversion what you expected? Are there any spurious characters? Do all the proteins begin with M?

Solution

- ☐ use UGENE to examine/alter start of this gene
- ☐ manually change start in Notepad/Notepad++

BLASTp as an annotation tool

- ☐ run your proteins against Caudoviricetes (taxid:2731619)
- ☐ always compare size of your protein with the consensus of its NCBI “hits”
- ☐ don’t concentrate on first “hit” since it may say “DNA polymerase” while all the rest say “hypothetical protein”
- ☐ Geneious has a very useful BLAST tool; more user friendly than NCBI BLAST

What do I call my gene product (i.e. protein)?

- ❑ “phage hypothetical protein” – redundant, all these proteins are “phage proteins”
→ hypothetical protein
- ❑ “gp87” (gp = gene product)
 - gp200 describes radically different proteins in *Listeria*, *Enterococcus*, *Mycobacterium*, *Rhodococcus*, *Sphingomonas*, *Pseudomonas*, *Bacillus* and *Synechococcus* phage genomes
 - Add /note=“similar to gp43 of Escherichia phage T4”

Gene Product Nomenclature 2

- ❑ /product=“UboA”; “Mcp”; “NrdA”; “hypothetical protein SA5_0153/152”; “ORF184” (as bad as gp184); “RNAP1”; “32 kDa protein”
 - Bad because they don’t mean anything to the casual (or informed) reader.
- ❑ **N.B.** Unless you are a bioinformatician or biostatistician be very conservative in recording “hits.” Could you convince your grandmother?, if not, list as a “hypothetical protein” but do take a stand -“putative DNA polymerase” is cowardly
- ❑ “Grandmother rule”



Use Consistent Nomenclature

- ❑ All of these describe homologs of the products related to the coliphage T4 *rIIA* protein!

rIIA protector from prophage-induced early lysis
 protector from prophage-induced early lysis
 protector from prophage-induced early lysis rIIA
 membrane-associated affects host membrane ATPase
 rIIA membrane-associated affects host membrane ATPase
 phage rIIA lysis inhibitor
 rIIA protector
 rIIA
 rIIA protein
 membrane integrity protector
 hypothetical protein
 unnamed protein product
 protein of unknown function

!!!!!!

Nomenclature Sins

- ❑ hypothetical protein → DNA polymerase with no or poor quality evidence is far worse than:
DNA polymerase → hypothetical protein
- ❑ Miss-annotation creep → database poisoning
- ❑ Be cautious about employing BLASTp hits in naming proteins – is there additional evidence to back the designation up?

Resources

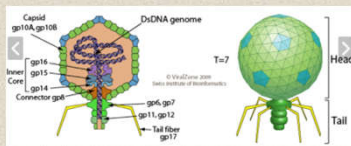
- ❑ UniProt Knowledgebase (UniProtKB) is a catalog of information on proteins with is manually curated and reviewed (check **Proteomes**). (<https://www.uniprot.org/>). Includes a BLAST feature.

Entry	Entry name	Protein names	Gene names	Organism
P00806	ENLYS_BPT7	Endolysin	3.5	Enterobacteria phage T7
P00581	DPOL_BPT7	DNA-directed DNA polymerase	5	Enterobacteria phage T7
P03696	DNBI_BPT7	Single-stranded DNA-binding protein...	2.5	Enterobacteria phage T7
P03726	EXLYS_BPT7	Peptidoglycan transglycosylase	16	Enterobacteria phage T7
P00638	EXRN_BPT7	Exonuclease	6	Enterobacteria phage T7
P00969	DNLI_BPT7	DNA ligase	1.3	Enterobacteria phage T7
P00641	ENDO_BPT7	Endonuclease I	3	Enterobacteria phage T7
P19726	CAPSA_BPT7	Major capsid protein	10	Enterobacteria phage T7

e.g. “capsid protein” versus head protein

Resources 2

- ❑ ViralZone (<https://viralzone.expasy.org/>) - a knowledge resource to understand virus diversity. Click on proteome for any viral genus.
- ❑ Linked to UniProt Knowledgebase (UniProtKB)



Extending the annotation

- ☐ using the phage_proteins.faa file
- ☐ general protein motif searches:

 Batch Web CD-Search Tool

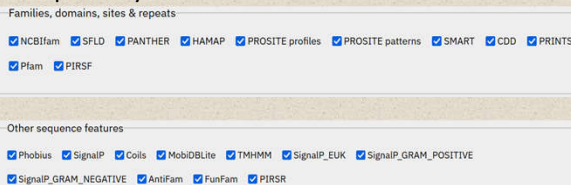
- Batch Web CD-Search Tool - <https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>
Change "Expect value" to 0.00001
- InterPro - <https://www.ebi.ac.uk/interpro/>



- Open latter results (*.tsv) in Excel

InterPro

- ☐ Under "Advanced options" consider running separately:



Transmembrane domains

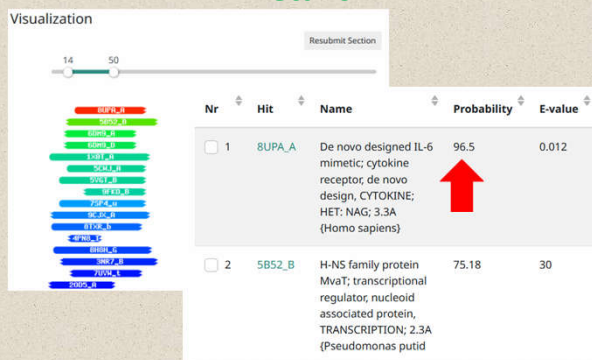
- ☐ using the protein.faa file
- ☐ a good resource:
 - DeepTMHMM - <https://services.healthtech.dtu.dk/services/DeepTMHMM-1.0/>
- ☐ in the case of Jeff six proteins, all identified as "hypothetical proteins," were identified to possess transmembrane domains (TMDs)
- ☐ in *.gbk file change "hypothetical protein" to "hypothetical membrane protein" and add: /note="three transmembrane domains identified using DeepTMHMM"



HHpred

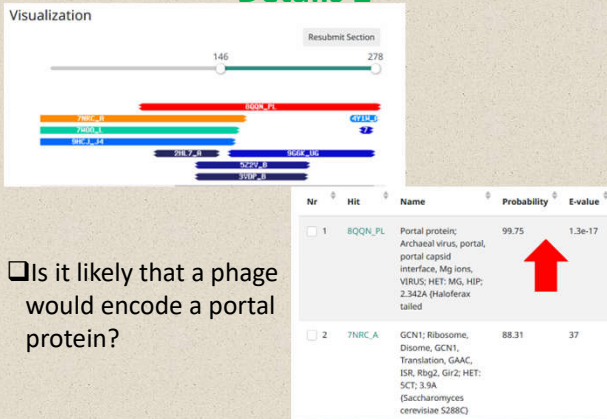
- ☐ Identification based upon similar structure
<https://toolkit.tuebingen.mpg.de/tools/hhpred>
- ☐ Introduction - <https://seaphages.org/video/87/>
- ☐ HHpred is useful for remote protein homology detection and structure prediction
- ☐ cannot be run remotely in batch mode
- ☐ worth local installation if you are seriously into phages

Details



- ☐ Is it likely that a phage would encode a cytokine?

Details 2



- ☐ Is it likely that a phage would encode a portal protein?

Phage therapy important questions

- ☐ Is my phage temperate or lytic?
 - ☐ Does it carry virulence determinants?
 - ☐ Does it carry antimicrobial resistance genes?
- The latter two points are dealt with by DFAST
- Alternative approaches listed next

Phage therapy important questions

- ☐ Is my phage temperate or lytic?
- Phage Lifestyle Prediction Servers



Phage AI

- <https://www.phage.ai/>
- requires registration

PhaBOX

- <https://phage.ee.cityu.edu.hk/phabox>
- PhageGE – Lifestyle prediction
- https://jason-zhao.shinyapps.io/PhageGE_Update/



AMR & Virulence Factors

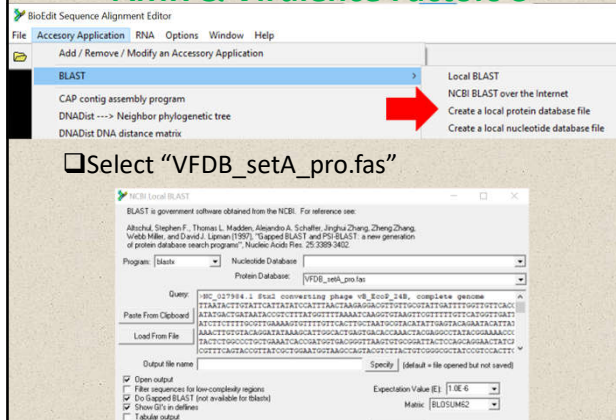
- ☐ Antimicrobial resistance – CARD
(Comprehensive Antibiotic Resistance Database) - <https://card.mcmaster.ca/>
- Use RGI Resistance Gene Identifier with DNA sequence at:
<https://card.mcmaster.ca/analyze/rgi>

AMR & Virulence Factors 2

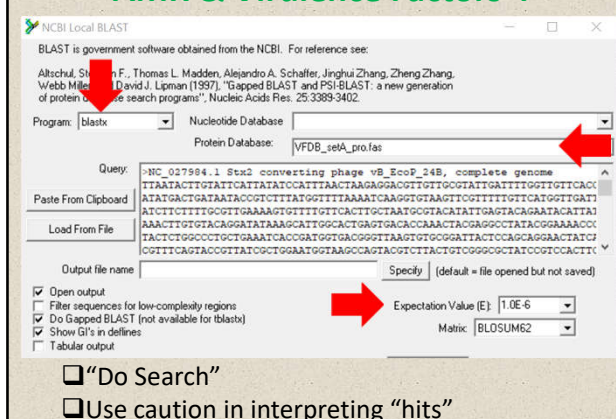
- ❑ Virulence factors can be found using VFAnalyzer at VFDB (Virulence factors of Pathogenic Bacteria)
 - <https://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi>
 - SLOW
- ❑ OR: Local BLAST versus VFDB “Protein sequences of core dataset)” using BioEdit 7.7 (Windows)
 - <https://bioedit.software.informer.com/>
 - OR UGENE



AMR & Virulence Factors 3



AMR & Virulence Factors 4



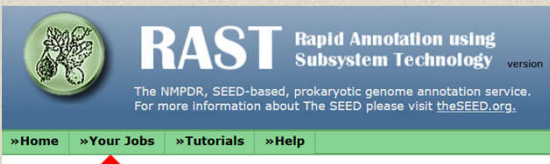
End of Part 2



Questions?

Notes on RAST

Create new project in RAST



- ☐ Choose this
- ☐ Select "Upload New Job"

RAST Step 2

File Upload:

Sequences File Jeff.fasta

Use this data and go to step 2

Contig statistics

Statistic	As uploaded	After splitting into scaffolds
Sequence size	66776	66776
Number of contigs	1	1
GC content (%)	55.6	55.6
Shortest contig size	66776	66776
Median sequence size	66776	66776
Mean sequence size	66776.0	66776.0
Longest contig size	66776	66776
N50 value		
L50 value	1	1

RAST Step 3

Genome Information:

Taxonomy ID:

Use this number for Caudoviricetes

Taxonomy string:

Domain: ☐ Bacteria ☐ Archaea ☒ Viruses

Genus: Insert Virus here

Species:

Strain:

Genetic Code: ☒ 11 (Archaea, most Bacteria, most Viri, and some Mitochondria)
☐ 4 (Mycoplasmata, Spiroplasmata, Ureoplasmata, and Fungal Mitochondria)

Use this data and go to step 3

RAST Step 4

Upload a Genome

Complete Upload

Please consider the following options for the RAST annotation pipeline

RAST Annotation Settings:

Choose RAST annotation scheme

Customize RASTtk pipeline ☐ Yes

Automatically fix errors? ☒ Yes

Fix frameshifts? ☐ Yes

Build metabolic model? ☐ Yes

Compute similarity? ☐ Yes

Turn on debug? ☐ Yes

Set verbose level

Disable replication ☐ Yes

Finish the upload

RAST – waiting for results

RAST Rapid Annotation using Subsystem Technology

The NHPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit [theSEED.org](#).

»Home »Your Jobs »Tutorials »Help

Your upload will be processed as job **1588956**. [View job status](#)

Go back to the [genome upload page](#) to add another annotation job.
You can view the status of your project on the [status page](#).

Job Details #1588956

» Available downloads for this job: RASTtk workflow ▾ | Download | Update download files

» [Share this genome with selected users](#)

» [Back to the Jobs Overview](#)

✓ Genome Upload has been successfully completed.

Genome ID - Name: 2731619.784 - Virus Pseudomonas phage Jeff

Job: #1588956

User: kropinski

Date: Thu Feb 20 09:40:01 2025

Genetic code: 11

Annotation scheme: RASTtk

Preserve gene calls: no

Automatically fix errors: yes

Fix frameshifts: no

Backfill gaps: yes

Check periodically since this page does not refresh

AK1

RAST Jobs Overview

Job ▾	Owner ▾	ID ▾	Name ▾	Num contigs ▾	Size (bp) ▾	Creation Date	Annotation Progress	Status
1588956	Kropinski, Andrew	2731619.784	Virus Pseudomonas phage Jeff	1	66776	2025-02-20 09:40:01	[view details]	complete

» Browse annotated genome in SEED Viewer

» Available downloads for this job: Genbank ▾ | Download | Update download files

» [Share this genome with selected users](#)

» [View Close Strains for this job](#)

» [Back to the Jobs Overview](#)

✓ Genome Upload has been successful

Genome ID - Name: 27311

Job: #158

User: kropi

Date: Thu F

Genetic code: 11

Annotation scheme: RAST

Preserve gene calls: no

Automatically fix errors: yes

Fix frameshifts: no

Backfill gaps: yes

✓ Rapid Propagation has been successful

✓ Quality Check has been successfully completed

For detailed explanations of the terms used in our quality report, please refer to [our wiki](#).

Annotations: Amino-Acid FASTA file, DNA Contigs, EMBL, EMBL (EC numbers stripped), GFF3, GFF3 (EC numbers stripped), Genbank, Genbank (EC numbers stripped), Genbank merged (for use in e.g. Artemis), Genome Typed Object (GTO), Genome directory, Nucleic-Acid FASTA file, RASTtk workflow, Spreadsheet (Excel XLS format), Spreadsheet (tab-separated text format)

Cleaning up RAST output

```
CDS      9476..11773
/db_xref="SEED:fig|2731619.784.peg.21"
/translation="MFKLWIFGRKKDSAAACESAPEKVAQIPDHDPIMIKLGRIR
GWNVEPEKAFVIRSVKDFLEPGLSVAMD SAYGDPTPAKVAAGGNFYVPTMLQDW
YNSQGFIGHQACALISQHWLVDKACMSGEDAARNWELKSDGRKLSDEQSALIRRD
MEFRVKNLVELNRFKNVFGVRIALFVVESDDFYYEKFNFEDGITPGSYKGISQVD
YMANQLTAASRADPSAEHFYEPDFWILSGKKYRSHLVVVRGPQPPDILKPTLFGG
TIPLTGRIVRYVVAERTANEAPLAMSRTSTIRHVDKALANEDAFNARLAFVIANR
DNHGVKVLIDEGMEQFTNLADFDISIIMNQYQLVAAIAKTATKLLGTSRPGFNATG
EHETISYHEELESIQEHIFDPLLRHYLLAKSEIDVQLEIVWNPVDSSTSSQQQAE
NNKGAATDEIYINSGVSPDEVRELRDDPRSGYNRLTDDQAEPEGMSPENLAEFEK
AGASVKAKEAEAEAAQAGAVEGAGGVPAAPRATKPLAKAEAGASAEPEPSPRP
PFAELRLKLVLLSKGLDIDIKAPGVVDIEHDAQGLAKTSKPOVSNPEPSPSNR
TVGPDHSELQRIKVNQITLTENPFGSIRQGGQSWKVMQMSHYGFTKTKGADGDE
VDCFVGNLGRKRVFVNVGNKEQDFDEHKMLGFNNINDAKSGYLSCFRPPGMDGLGS
IHEVDLPAFRRWLANGDTTKFFGGG
/product="Phage protein"
/transl_table=11
11773..12609
/db_xref="SEED:fig|2731619.784.peg.22"
/translation="MAFKASKKRERRAPLFVGRGKPIIPSAGIEAWYRQMKMDSKIM
ISDYRNEIEKALSQPAERFFARDESNNVLFKMTLSLQQRSRIFEQFAAKIAPEFY
NRTEAATAATLHSLSVAGVDQPRAAVNSVNTLEAATTNHTLITKIQEEVHEKIY
TSVMSLSTSFNEEQGTSGITNALRKVKGFSEDRILARDQTSKLYSLSDERMAEN
GVSEFEWLHSSAGKTPPHLHLEKDKGRKFLNDRLEWGGFRADQPPGFWAINCRKRI
PVI"
/product="Phage minor capsid protein"
/transl_table=11
```

No locus tags

Phage is redundant

Slide 65

AK1 Andrew Kropinski, 2025-02-20

Clean-up using Notepad++

- ☐ Under “Search” and “Replace”

Replace

Find what: `cb_xref=SEED:fgj2731619.784.png`

Replace with: `locus_tag=J3EFF_`

☐ In selection

☐ Backward direction
☐ Match whole word only
☐ Match case
☐ Wrap around

Search Mode

☒ Normal
☐ Extended (n, \r, \t, \0, \x...)
☐ Regular expression ☐ matches newline

☒ Transparency
☒ On losing focus
☐ Always

Find Next
 Replace
 Replace All
 Replace All in All Opened Documents
 Close

- ❑ Under “Search” and “Find” modify “Phage”

Final version

```

CDS      9476..11773
         /locus_tag="JEFF_21"
         /translation="MFKLSVIFGRKKDSAAACESAFKVAQIQHPLDPMIKLGRW
         GWNVEFEKAPVIRSVKQFLPEGLSVAMDSAVSDGGTFAAKVAAGGQNVVPTMLQDRI
         YNSQGFIGHCAQALISQHWLVKAKMSGDFGAAANRHLSDGKRLSDGSLALAIRL
         MEYFVKVNLVLRNVPVQVRLALFWLSDGQVETKFFPFGISITQSPKGISQVDI
         YWAMPLQSTADTASDEAPSRVIVLISGKYHNSHLVYVQSPQDPLKTLPTFLFG
         IPLTQIRVYVAERATNEAPFLMAMSKSTITIHVDKAEADENAFNARLFIANR
         DNHWGVKLVGDEMGQFTNLADSDSIHQNVQVLAIAKATPATKFLNSPKQMLGR
         EHETISYHEELSEIQEIRIFDLHYRLLYALKSEIDVQGLEWNPVDSSTQSAQAEK
         NKKGAATVIVINSQVSYFEVERLRDQDPKSGNLIITQVNEFPGSGSPNIAEEL
         AGQGVQKAEKAEKQKNGVFGGFGFADPPTKFLALAAEGASAEALVQV
         PKAEALNLDVLLSKQLQDQDKQFQVQIDHNAPDQKSTKSPGVSGSEPSVFSNR
         TVGDFRQKQIKYKNGVNTLTIEENPFGISRGKQSGKRWQKGVFKTKTGADGGE
         VDCFVGNLHSGRFRVFNQVKNKEQGFDEHKMKLGFNINDAKSGYLSCFRPGWDGLGS
         IHEVDFLWPKRILANGQDTKPFGE"
         /product="hypothetical protein"
         /trans_table=11
         11773..12609
         /locus_tag="JEFF_22"
         /translation="MAFYASKKRERRAPLVPRGQKIIPSAGIEIANYRKQMKMSKLI
         NSEYEATEIKALSGPAERFFFADESIVNLFQMTLSESLRSGRFEFGAFKAPFEV
         NRTEAANATLHLSVAGVDQPPAAINYSVNTLEAATTNHTLTIKQIEVKEHIY
         TVSMVLSLTFNPFEGQSTITNARKVYGRSEDLIELARQDTSLSLSDERWAM
         GVSEEFHSSAGKTPHRLHKEGKFKLMDPLWEGKAGQFGPWAINCRKPT
         VI"
         /product="minor capsid protein"
         /trans_table=11

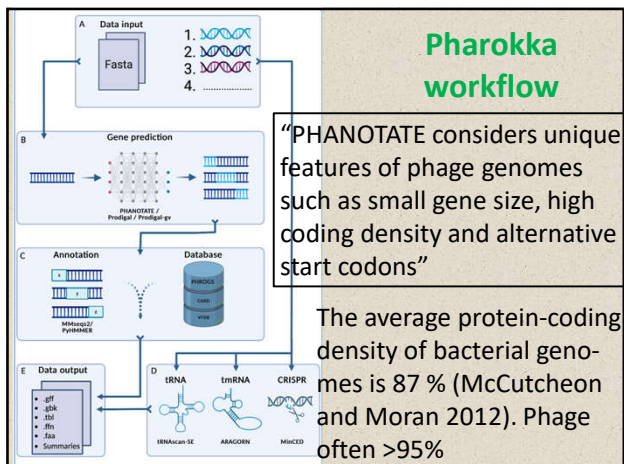
```

Notes on Pharokka

Pharokka – the only phage-specific pipeline

Galaxy
EUROPE



Pharokka via Galaxy

Galaxy
EUROPE

- ☐ <https://usegalaxy.eu/>
- ☐ Also available at: <https://usegalaxy.org.au/>
- ☐ Need to register to use these servers

Start here



Uploading genome sequence

Upload from Disk or Web to **Unnamed history**

Regular Composite Collection Rule-based

Drop files here

Type (set all): Auto-detect Reference (set all): unspecified (?)

Choose local file Choose remote files Paste/Fetch data Start Pause Reset Close

History

search datasets

Unnamed history

8.32 kB

4: JEFF FOR ANNOTATION.fasta

Choose Tools

Galaxy Europe

Upload Tools Workflows

Tools

search tools

Get Data Send Data Collection Operations

Tools

Pharokka

Show Sections

Pharmacophore generation (Align-It)

Pharmacophore alignment and optimization (Align-It)

pharokka Rapid standardised annotation tool for bacteriophage genomes and metagenomes

pharmCAT Pharmacogenomics Clinical Annotation Tool

Choose parameters & Run

pharokka Rapid standardised annotation tool for bacteriophage genomes and metagenomes (Galaxy Version 1.3.2+galaxy9)

Pharokka DB version 1.2.9 downloaded at 2023-08-27 07:02:08.916437

Using built-in pharokka DB

User specified gene predictor *

Phanotate

Meta mode for metavirome input samples

No

E-value threshold for mmseqs2 PHROGs database search. Defaults to 1E-05. *

1e-05

Runs - terminase large subunit - re-orientation mode. Single genome input only and requires terminase_start to be specified.

Do not run 'terminase large subunit' re-orientation mode.

Create a Zip archive of the complete output for further investigation.

Yes

Run Tool

Running and output



History

search datasets

Unnamed history

899 kB

7: pharokka on data 4: zip of the complete output

6: pharokka on data 4: GFF

5: pharokka on data 4: Genbank

4: JEFF FOR ANNOTATION.fasta

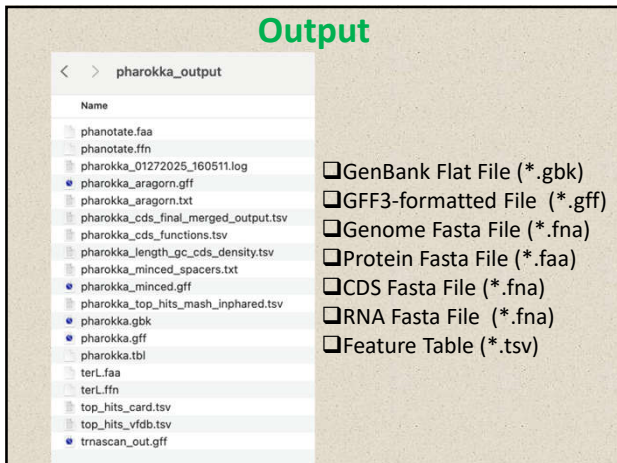
Started tool pharokka and successfully added 1 job to the queue.

It produces 3 outputs:

- 5: pharokka on data 4: Genbank
- 6: pharokka on data 4: GFF
- 7: pharokka on data 4: zip of the complete output

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Output



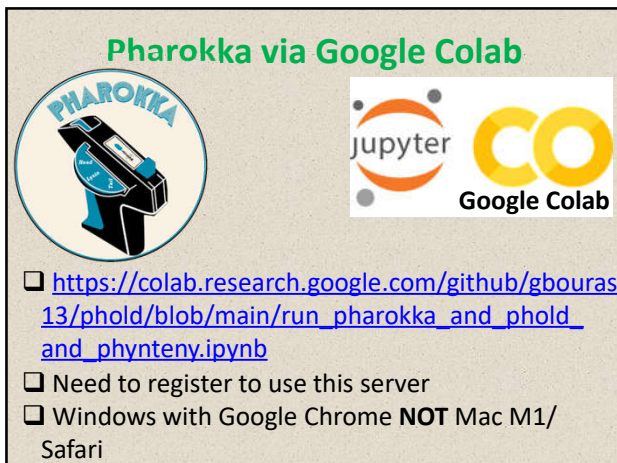
< > pharokka_output

Name

- phanotate.faa
- phanotate.ffn
- pharokka_01272025_160511.log
- pharokka_aragorn.gff
- pharokka_aragorn.txt
- pharokka_cds_final_merged_output.tsv
- pharokka_cds_functions.tsv
- pharokka_length_gc_cds_density.tsv
- pharokka_minced_spacers.txt
- pharokka_minced.gff
- pharokka_top_hits_mash_inphared.tsv
- pharokka.gbkl
- pharokka.gff
- pharokka.tbl
- terL.faa
- terL.ffn
- top_hits_card.tsv
- top_hits_vfdb.tsv
- trnscan_out.gff

- ☐ GenBank Flat File (*.gbk)
- ☐ GFF3-formatted File (*.gff)
- ☐ Genome Fasta File (*.fna)
- ☐ Protein Fasta File (*.faa)
- ☐ CDS Fasta File (*.fna)
- ☐ RNA Fasta File (*.fna)
- ☐ Feature Table (*.tsv)

Pharokka via Google Colab



https://colab.research.google.com/github/gbouras13/phold/blob/main/run_pharokka_and_phold_and_phynteny.ipynb

- ☐ Need to register to use this server
- ☐ Windows with Google Chrome **NOT** Mac M1/ Safari

Advantages of this version over Galaxy version

- ❑ Allows one to specify the Locus tag name – use name of phage
- ❑ Automatically generates zipped file

Loading programs and databases

run_pharokka_and_phold_and_phynteny.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Pharokka + Phold + Phynteny

pharokka is a rapid standardised annotation tool for bacteriophage genomes and metagenomes. You can read more about pharokka in the [documentation](#).

phold is a sensitive annotation tool for bacteriophage genomes and metagenomes using protein structural homology. You can read more about phold in the [documentation](#).

1. Install pharokka and phold

```
#%%title 1. Install pharokka and phold
#markdown This cell installs pharokka and phold. It will tel
%%bash
```

2. Download pharokka phold databases

```
#%%title 2. Download pharokka phold databases
#markdown This cell downloads the pharokka then the phold d
```

Loading sequence

2. Upload sequence

Files

pharokka_db

phold_db

sample_data

CONDA_READY

PHAROKKA_PHOLD_READY

3. Run Pharokka

INPUT_FILE: Jeff.fasta

PHAROKKA_OUT_DIR: output_pharol

GENE_PREDICTOR: phanotate

PHAROKKA_PREFIX: pharokka

LOCUS_TAG: Jeff

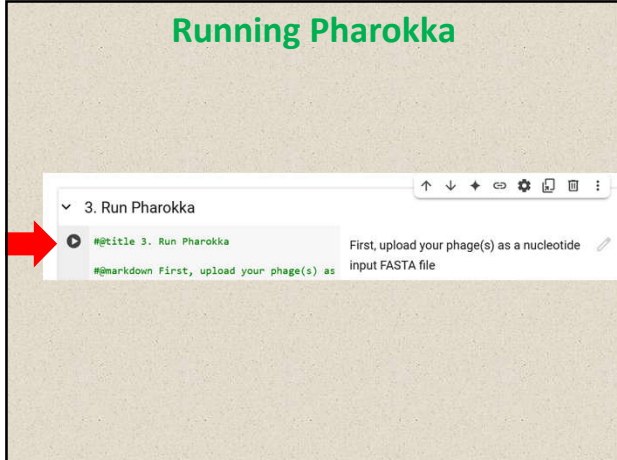
FAST: ☒

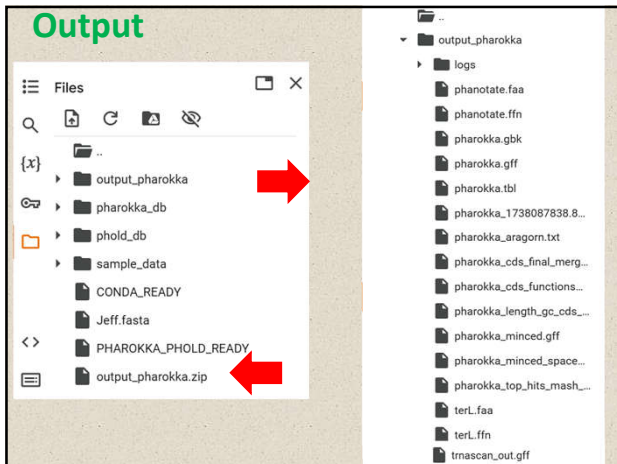
1. Choose file (*.fasta)

3. Same name

4. Adjust

Running Pharokka





Problems with Pharokka

- ☐ locus_tags include the word CDS and are sequential from 0001 – N.B. phages don't have thousands of genes
- ☐ locus_tag numbers increase by one unit – N.B. if a CDS is missing you will have to use A, B, C after the number.
- ☐ translation code may not be included in *.gbk file.
- ☐ you cannot change the translation code (11).
N.B. Some large, uncultivated phages of the gut microbiome - predominantly Lak phages and crAssphages - have recoded the TAG or TGA stop codon (genetic codes 15 and 4).

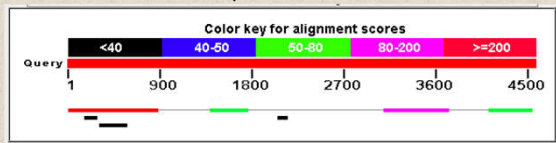
Problems with Pharokka 2

- ☐ annotation is **not perfect**
- ☐ splicing not recognized
- ☐ in GenBank record this is reported as CDS
join(99941..100807,101606..103957)
- ☐ important with members of the
Herelleviridae, such as *Staphylococcus* phages;
and, *Campylobacter* phages

Introns and Inteins – rare but

- ☐ the gene encoding the aerobic ribonucleoside diphosphate reductase (large subunit) from *Campylobacter* phage vB_CcoM-IBB_35 is located on a 4.5 kb region which homologous to a 2.3 kb region from *Klebsiella* phage vB_KleM-RaK2. N.B. proteins have similar mass.

N.B. the thin grey line joining blocks indicates a downstream sequence which is similar

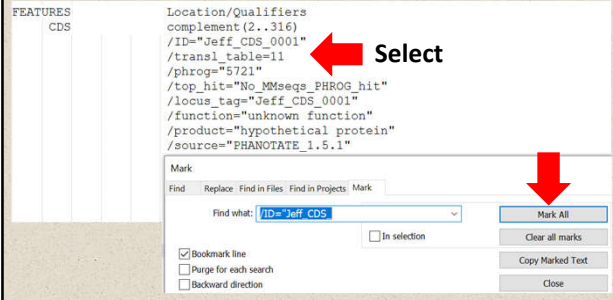


Editing primary annotation (*.gbk) files

- ☐ No matter what program you use you will have to edit the results
- ☐ Notepad++ is better than Notepad

Open in Notepad++

- ❑ Install from: <https://notepad-plus-plus.org/>
- ❑ Review: <https://www.youtube.com/watch?v=wVfbFe57h2o>
- ❑ Under "Search" choose "Mark" and "Mark All"



Mark All/Delete

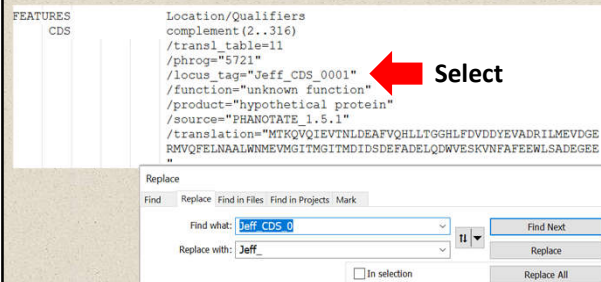
Location/Qualifiers
complement (2..316)
/ID=Jeff_CDS_0001"
/transl_table=11
/phrog="5721"
/top_hit="No MMseqs_PHROG_hit"
/locus_tag="Jeff_CDS_0001"
/function="unknown function"
/product="hypothetical protein"
/source="PHANOTATE_1.5.1"
/score="-56.95485423688936"
/phase="0"
/translation="MTKQVQIEVTNLDEAFVQHLLTGGLFDVDDYEVADRILMEVDGE
RMVQFELNAALWNMEVMTGTMGIDTMDISDEFADQLQDWVESKVNFAFEWLSDAGEEE
"

- ❑ Under "Search", choose "Bookmark" and "Remove Bookmarked Lines"

Location/Qualifiers
complement (2..316)
/transl_table=11
/phrog="5721"
/locus_tag="Jeff_CDS_0001"
/function="unknown function"
/product="hypothetical protein"
/source="PHANOTATE_1.5.1"
/translation="MTKQVQIEVTNLDEAFVQHLLTGGLFDVDDYEVADRILMEVDGE
RMVQFELNAALWNMEVMTGTMGIDTMDISDEFADQLQDWVESKVNFAFEWLSDAGEEE
"

Modify locus_tag

- ❑ Under "Search" choose "Replace" and "Replace All"



- ❑ Changes /locus_tag="Jeff_CDS_0001" to /locus_tag="Jeff_001"

Progress

FEATURES	Location/Qualifiers
CDS	complement (2..316) /transl_table=1 /phrog="5721" /locus_tag="Jeff_001" /function="unknown function" /product="hypothetical protein" /source="PHANOTAX.1.5.1" translation="MTKQVQIEVNLDEAFVQHLLTGGHLEFDDVDYVADRILMEVDGGE RMVQFLINAAALMNNVEVGITNGITMDIDSEFADELQDWVESKVNFVAFEEHLSADEGEE "
CDS	complement (65914..66776) /transl_table=1 /phrog="5834" /locus_tag="Jeff_101" /function="tail" /product="tail length tape measure protein" /source="PHANOTAX.1.5.1" translation="HGARFVSRAHRRIVSVSKLFFNSIDGQWFKLEVVDGAEIVLIVL VSVTSQGFKFWFVLELCEKAEKAEVFPVKSYSYLGATFENRALAPFAQGVLEI DQKRDDESEVSTDESITAMELDRENSRLDCEQLNSTAREKTAEVNQLQLRIQLEL EVESKNISATIGHQRRIDELLKHQSOMLSAQVVFSELESKLTLEALKQSLTMAKTI RLERAAKSATIDENKAETNSQLALRIRLEAELESQKALREEDRSRSLDGSITANANQI KLACEARLARSAREDAEFYRIGMQRIAEIANRKPFED"

Edit top of *.gbk file

	Pseudomonas	66776 bp	DNA	linear	PHG 28-JA
ON	Pseudomonas phage Jeff, complete genome.				
N	Pseudomonas				
	Pseudomonas				
	.				
	.				
SM	.				
	.				
	Location/Qualifiers				
	Jeff	66776 bp	DNA	linear	PHG 28-JAN-2025
ON	Pseudomonas phage Jeff, complete genome.				
N	Pseudomonas phage Jeff				
	Pseudomonas phage Jeff				
	.				
	.				
SM	Pseudomonas phage Jeff				
	Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviri				
	Pbnunavirus; Pseudomonas phage Jeff				
	Location/Qualifiers				

Editing DEFINITION, SOURCE and ORGANISM

- ❑ DEFINITION - Pseudomonas phage Jeff, complete genome
- ❑ SOURCE - Pseudomonas phage Jeff
- ❑ ORGANISM Pseudomonas phage Jeff
Viruses; Duplodnaviria; Heunggongvirae; Uroviricota; Caudoviricetes; Pbnavirus; Pseudomonas phage Jeff.

This will be discussed on the last day of this workshop
BUT in many cases the taxonomy can be readily
assessed from the BLASTn hits

