# phage annotation workshop

**Evelien Adriaenssens**

**Dann Turner**

**Andrew Kropinski**

# Who are we?

**Evelien Adriaenssens**

Group Leader, Quadram Institute Bioscience, Norwich, UK

Chair Bacterial Viruses Subcommittee ICTV

NCBI Genomes Advisor

**Dann Turner**

Lecturer, University of the West of England, Bristol, UK

Vice Chair Bacterial Viruses Subcommittee ICTV

*Caudoviricetes* Study Group Chair

**Andrew Kropinski**

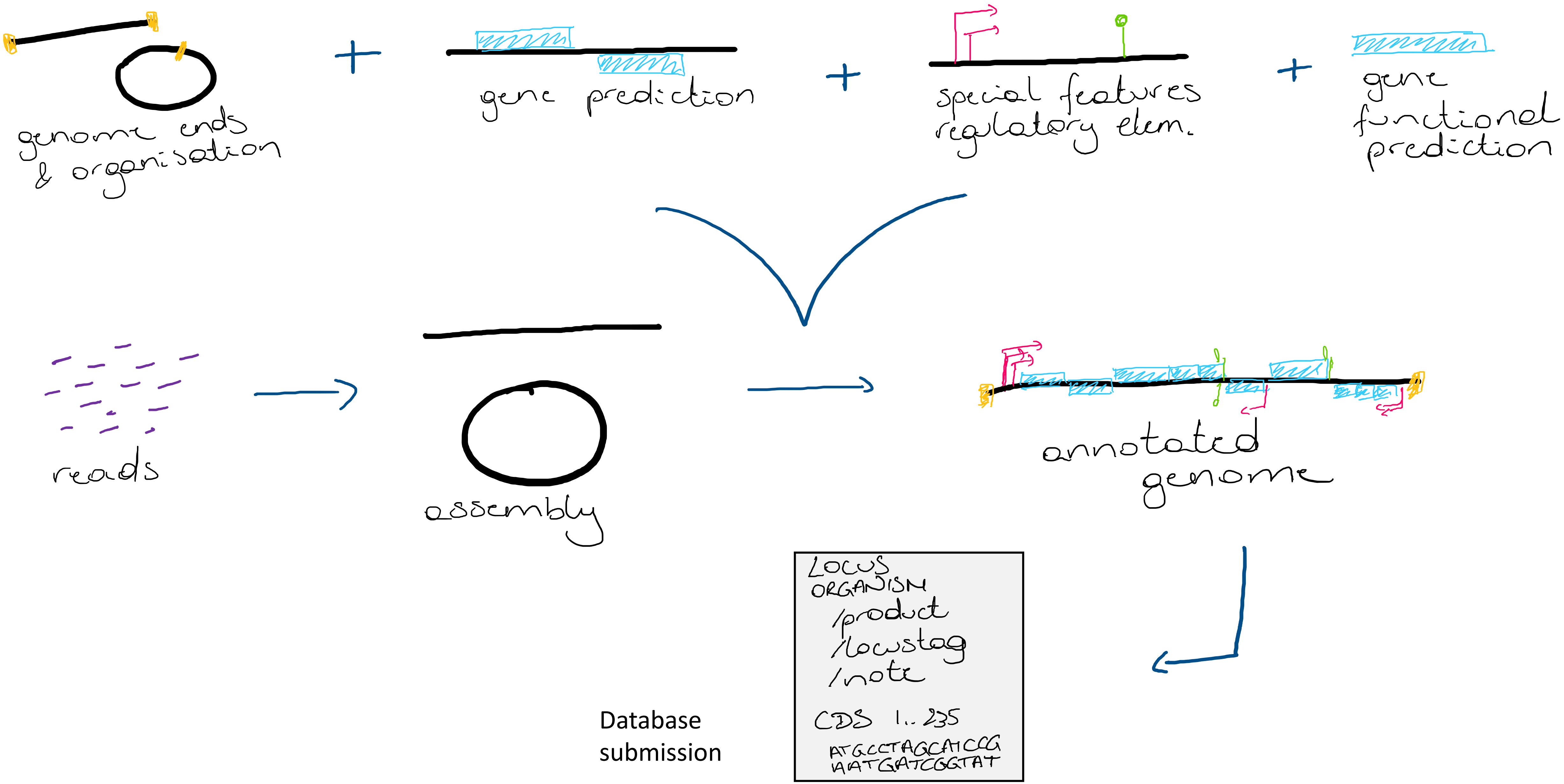
Emeritus professor, University of Guelph, Canada

former Chair Bacterial and Archaeal Viruses Subcommittee ICTV

NCBI Genomes Advisor

# Workshop overview

- Introduction
- Sequencing and assembly
- Genes in phage genomes (annotation)
- Intro to classification & taxonomy



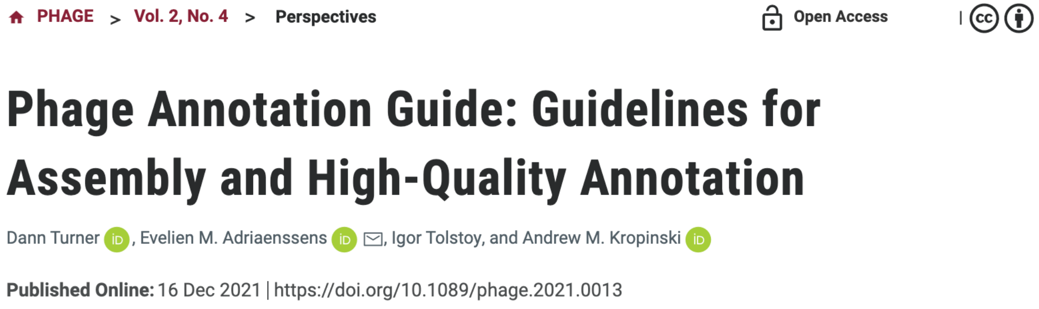
Database
submission

# Resources

- PHAGE journal Special Issue on Phage Informatics & AI
  - https://www.liebertpub.com/doi/10.1089/phage.2021.0013
  - https://www.liebertpub.com/doi/10.1089/phage.2021.0015

- Phage Annotation Workshop: QIB & AAFC Canada Partnership
  - https://github.com/quadram-institute-bioscience/phage-annotation-workshop/wiki

- Phage Annotation Workshop by Andy Millard (Sep 2022), contact Andy for more info

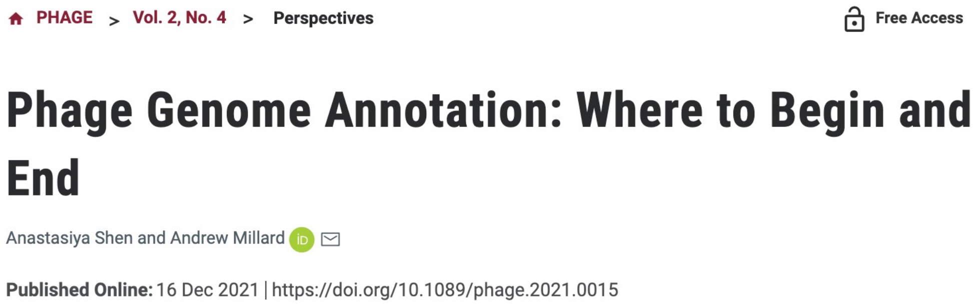


PHAGE > Vol. 2, No. 4 > Perspectives Open Access

## Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation

Dann Turner, Evelien M. Adriaenssens, Igor Tolstoy, and Andrew M. Kropinski

**Published Online:** 16 Dec 2021 | https://doi.org/10.1089/phage.2021.0013



PHAGE > Vol. 2, No. 4 > Perspectives Free Access

## Phage Genome Annotation: Where to Begin and End

Anastasiya Shen and Andrew Millard

**Published Online:** 16 Dec 2021 | https://doi.org/10.1089/phage.2021.0015

**phage** annotation workshop

# Phage Genome Sequencing and Assembly

Dann Turner (dann2.turner@uwe.ac.uk)

# Overview

- Sequencing and assembly
- Orientating phage genomes
- Frameshift errors
- Genome termini

# Errors in Submitted Sequences in 2022

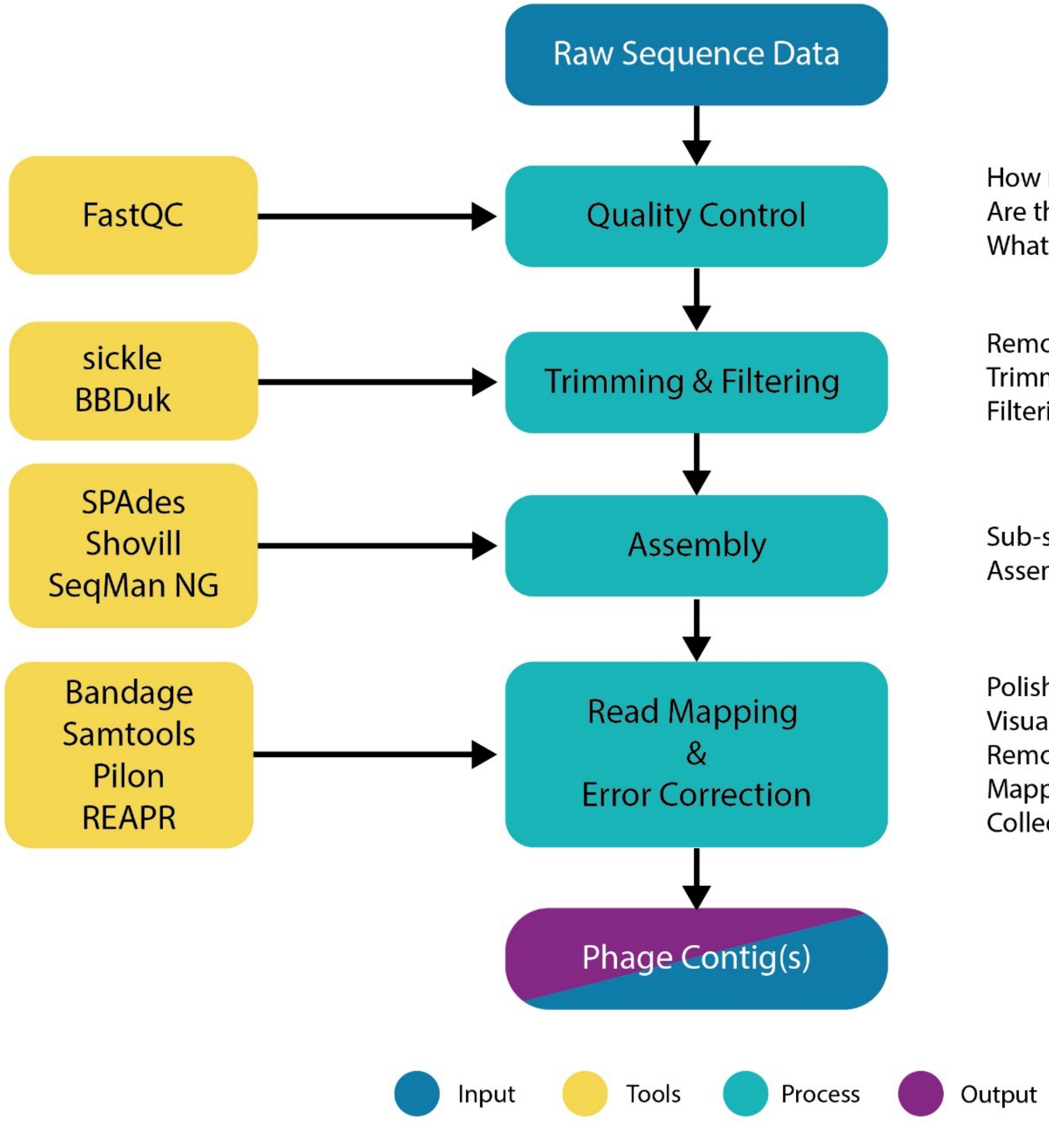
- Sequence errors (43%)
  - Frameshifts, genome too long or too short
- Incorrect taxonomy (29%)
  - TEM micrograph does not match sequence
  - Not identified as a prophage
  - Wrong host identified
- Chimeric genomes (21%)
  - Two phages, co-assembly of 16S rDNA, mitochondrial DNA present
- Duplicated or incorrect phage names (7%)
- Genome not colinear with type phage (very common)
- Genome identified as circular (very common)

# Sequencing Platforms

| Platform | Pros | Cons |
|---|---|---|
| **Illumina** | Lowest error rates | Long sequencing runs |
| | Widely used and range of instruments | Polymerase bias |
| | Lowest per-Gb cost | High instrument costs |
| | High output yield | |
| **PacBio** | Long reads | Low output yield |
| | Fast sequencing runs | High(ish) error rates |
| | Detection of base modifications | Massive instrument cost |
| **ONT** | Fast | High error rate |
| | Longest read length | Sensitivity of nanopores |
| | Low cost of instrument and consumables | Technical expertise required for data analysis |
| | Detection of base modifications | |

Adapted from Clin. Microbiol. Rev. 30(4):1015

# Sequencing and Assembly Overview



Raw Sequence Data

FastQC → Quality Control

sickle
BBDuk → Trimming & Filtering

SPAdes
Shovill
SeqMan NG → Assembly

Bandage
Samtools
Pilon
REAPR → Read Mapping & Error Correction

Phage Contig(s)

● Input ● Tools ● Process ● Output

How many reads are available?
Are there adaptors present?
What are the quality statistics like?

Removal of adaptors
Trimming of low-quality bases (5′ and 3′)
Filtering of low-quality reads

Sub-sample reads to 30-100x coverage
Assembly

Polishing for error correction
Visualise assembly graph with Bandage
Removal of small and very low coverage contigs
Mapping of reads back to assembled contigs
Collection of reads that do not map to the phage contig

# Library preparation and coverage

- Avoid library preparation kits that rely upon transposon-mediated shearing and adaptor ligation (e.g. NexteraXT)

- Use multiplexing to take advantage of HTS platform yield

- Remember that excessive coverage can be detrimental to assembly

- Coverage of ~100x is recommended

$$\text{number of reads} = \frac{(\text{coverage} \times \text{genome size (bp)})}{\text{read length (bp)}}$$

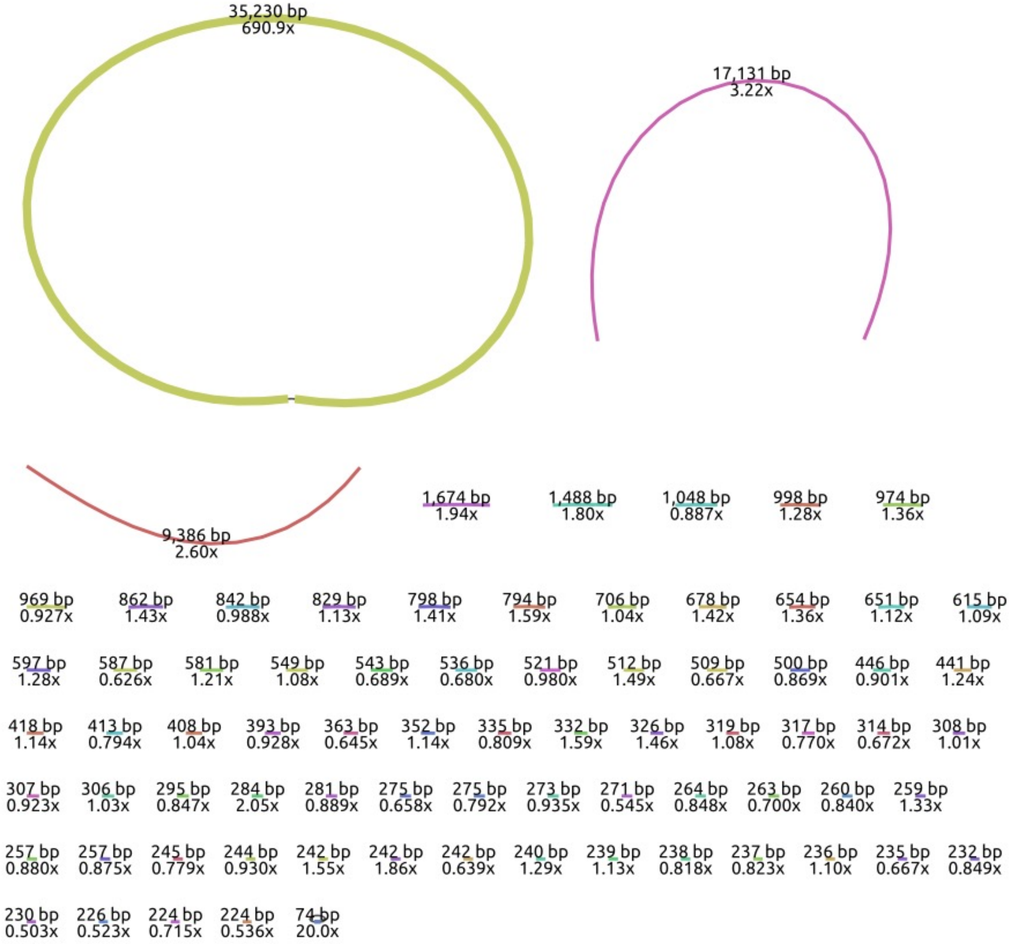
# Assembly

- **Short or long-reads**
  - SPAdes: for assessment see Rihtman et al., PeerJ 4:e2055
  - PacBio/ONT: Canu, Flye, Miniasm
  - Commercial GUI options: SeqMan NG/CLC Genomics

- **Hybrid assembly?**
  - Not really necessary for phage genomes (additional expense)
  - If using: short-read first vs long-read first (Unicycler and Trycycler)

# Assembly Validation

- **Bandage:** visualising the assembly graph
- **Mapping reads:**
  - Calculation of coverage
  - Identification of areas of low/high coverage
  - Identification of areas for targeted Sanger sequencing
  - Identification of reads not mapping to the phage contig – host DNA, prophages, mixed sample?
  - **QUAST, BWA-MEM, Bowtie2, Minimap2**



*A. baumannii* prophage assembly graph

# Troubleshooting

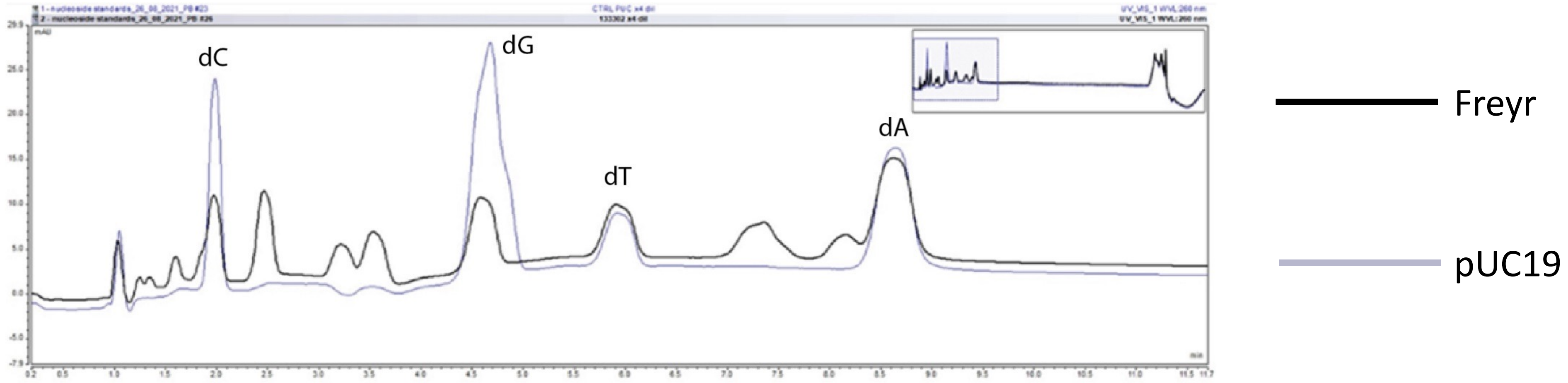
- An incomplete assembly can result from a number of factors
  1. Read coverage is excessive
  2. Mol G+C% bias
  3. Repeat sequences (e.g. IS elements)
  4. Presence of multiple similar phage genomes (high micro-diversity)

- Resolutions?
  1. Down sample number of reads before assembly
  2. PCR amplification method
  3. Normally only an issue when high amounts of background host DNA
  4. Mapping of reads

# Strategies for “hard to sequence” phages

- Some phages with hypermodified bases are refractory to traditional sequencing methods, e.g.
  - YerA41 (Viruses 2020;12:620)
  - Roseophages (Curr. Biol. 2021; 31:3199)

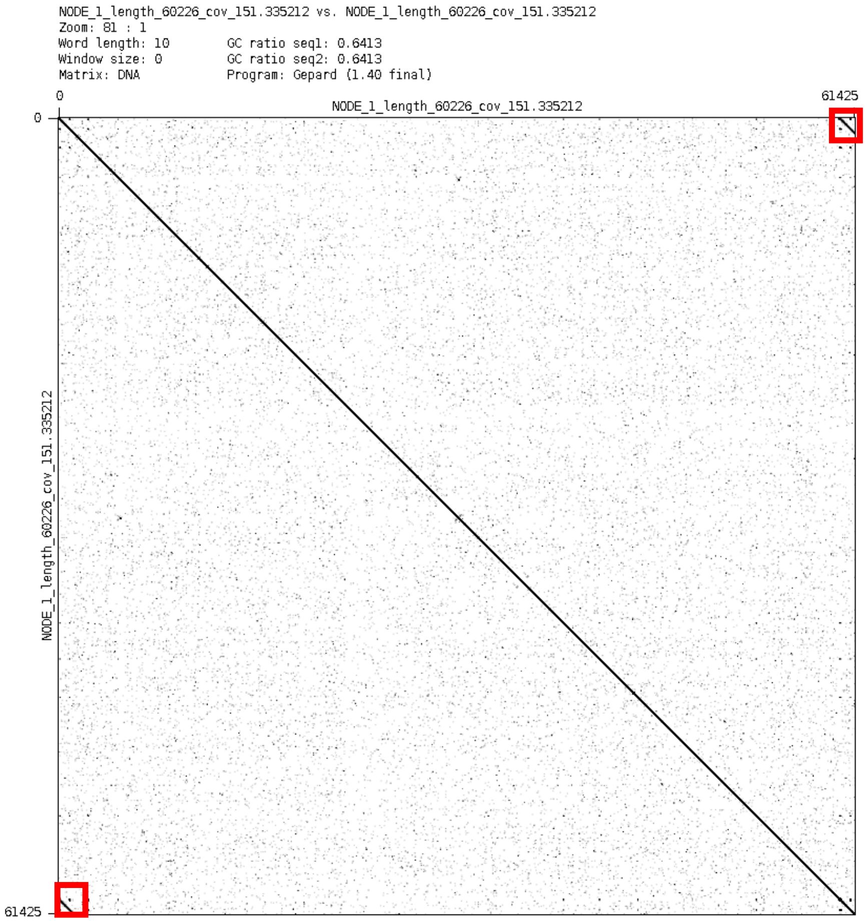
Freyr (13302)



- RNA-seq to reconstitute the genome from phage transcripts (expensive)
- Rolling circle amplification

# Orientating genomes

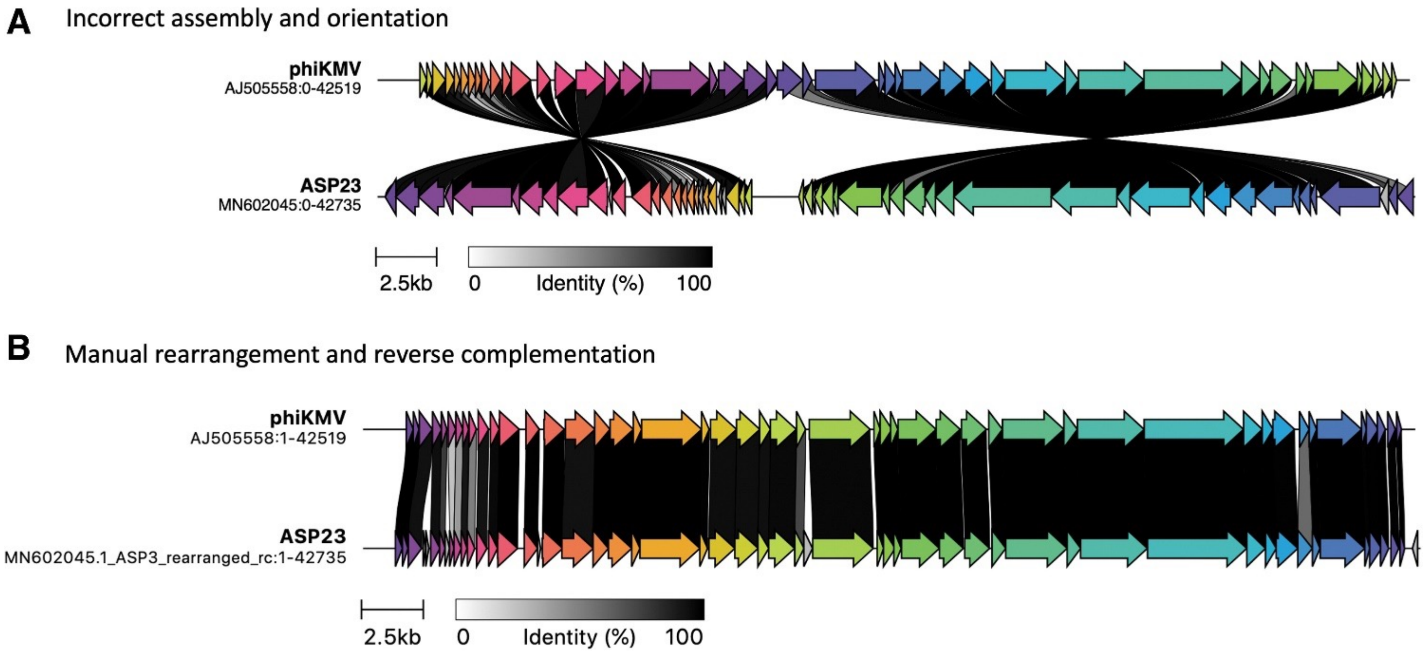
- Genomes of almost all known dsDNA phages are packaged as linear molecules
- Many assemblers will result in an apparently circular consensus contig
- Circularity is an artefact of the assembly process (but generally indicates a complete genome!)
- Reorientation may require reverse complementation and/or breaking and re-joining of the contig
- Important to assess genome termini first



```
NODE_1_length_60226_cov_151.335212 vs. NODE_1_length_60226_cov_151.335212
Zoom: 81 : 1
Word length: 10       GC ratio seq1: 0.6413
Window size: 0        GC ratio seq2: 0.6413
Matrix: DNA           Program: Gepard (1.40 final)
```

# Why Orientate?

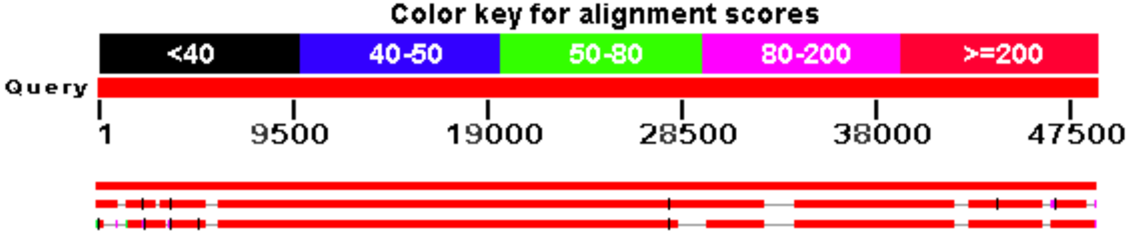
- Makes sequence comparisons more intuitive
- Allows for better pairwise visualisation (e.g. cLinker/EasyFig)


A
Incorrect assembly and orientation
phiKMV
AJ505558:0-42519
ASP23
MN602045:0-42735
2.5kb
0
Identity (%)
100
B
Manual rearrangement and reverse complementation
phiKMV
AJ505558:1-42519
ASP23
MN602045.1_ASP3_rearranged_rc:1-42735
2.5kb
0
Identity (%)
100

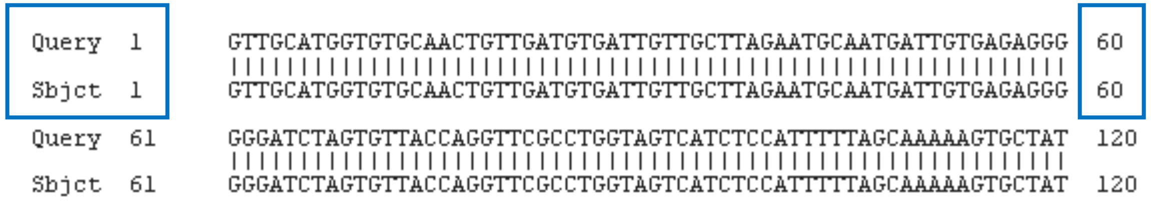
- Conventions
  - Orientate using genome termini (more on this next…)
  - Open at small or large terminase subunit (whichever is identifiable)
  - Open at rIIA gene (*Straboviridae*)

# Tools for orientation

- **BLASTn**
  - Phage vB_EcoP_AMK is closely related to three genomes

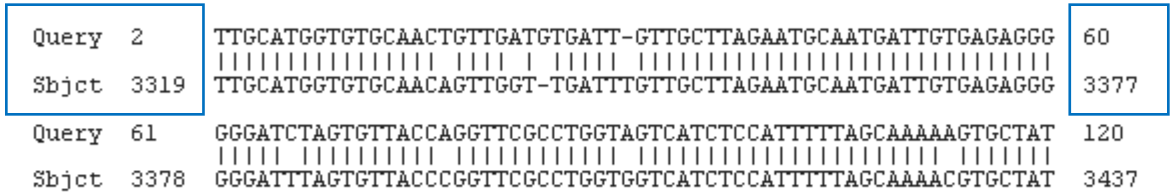


Colinear



```
Query  1     GTTGCATGGTGTGCAACTGTTGATGTGATTGTTGCTTAGAATGCAATGATTGTGAGAGGG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GTTGCATGGTGTGCAACTGTTGATGTGATTGTTGCTTAGAATGCAATGATTGTGAGAGGG  60

Query  61    GGGATCTAGTGTTACCAGGTTCGCCTGGTAGTCATCTCCATTTTTAGCAAAAAGTGCTAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GGGATCTAGTGTTACCAGGTTCGCCTGGTAGTCATCTCCATTTTTAGCAAAAAGTGCTAT  120
```

Not Colinear

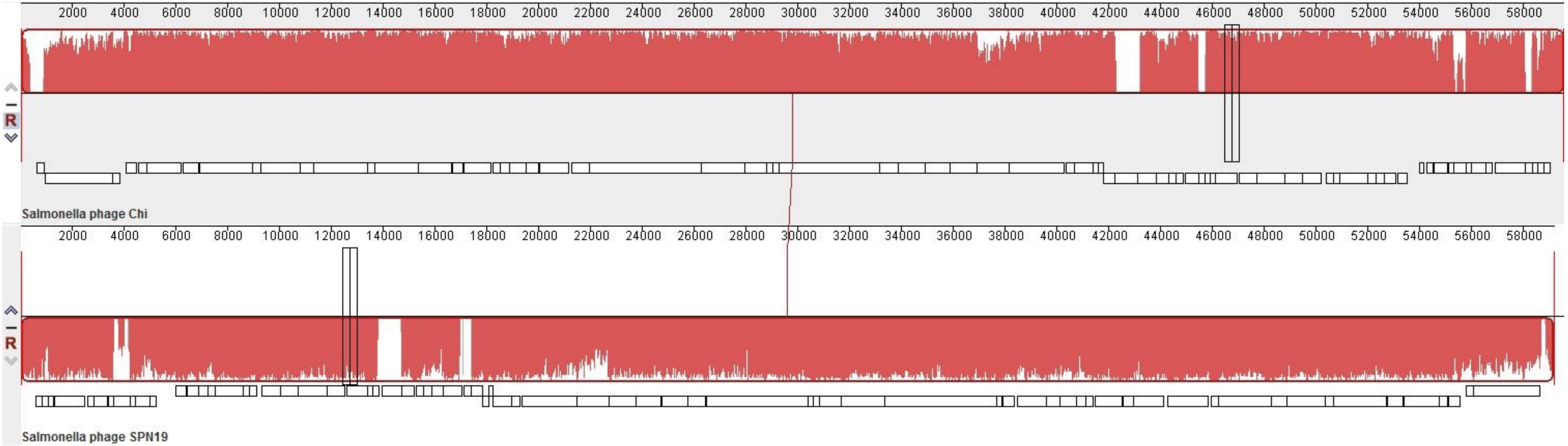


```
Query  2     TTGCATGGTGTGCAACTGTTGATGTGATT-GTTGCTTAGAATGCAATGATTGTGAGAGGG  60
             |||||||||||||||| |||| | ||||| ||||||||||||||||||||||||||||||
Sbjct  3319  TTGCATGGTGTGCAACAGTTGGT-TGATTTGTTGCTTAGAATGCAATGATTGTGAGAGGG  3377

Query  61    GGGATCTAGTGTTACCAGGTTCGCCTGGTAGTCATCTCCATTTTTAGCAAAAAGTGCTAT  120
             ||||| |||||||||| |||||||||||| |||||||||||||||||||||| |||||||
Sbjct  3378  GGGATTTAGTGTTACCCGGTTCGCCTGGTGGTCATCTCCATTTTTAGCAAAACGTGCTAT  3437
```

  - Limit searches to *Caudoviricetes* (taxid: 2731619) in Organism field

# Tools for orientation

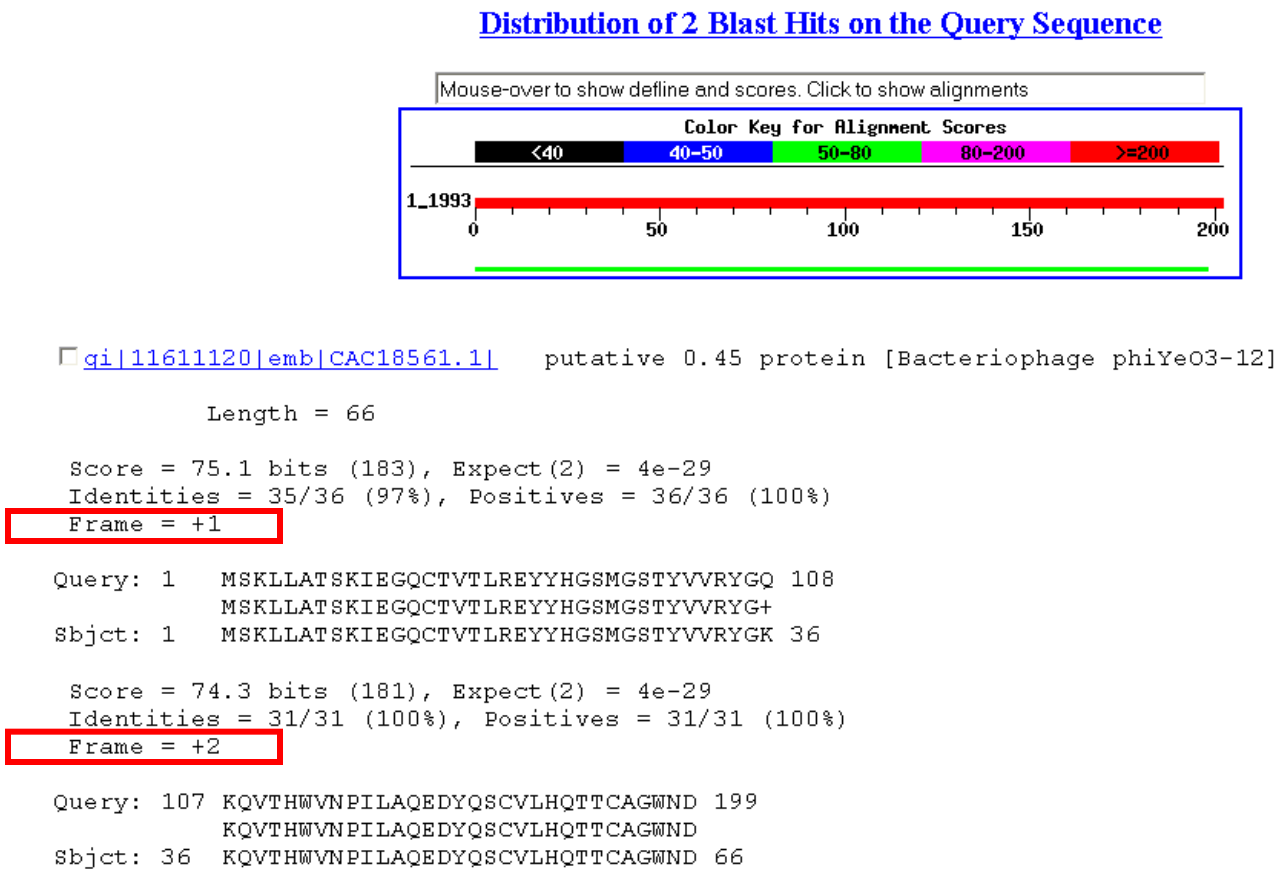
- **Progressive Mauve**
  - A bit problematic thanks to Java



- **Cutting, pasting and rejoining**
  - http://reverse-complement.com/
  - http://www.bioinformatics.org/sms/rev_comp.html
  - http://www.cellbiol.com/scripts/complement/dna_sequence_reverse_complement.php
  - https://notepad-plus-plus.org/downloads/

# Frameshifts

- BLASTx can be used to identify potential frameshifts if similar phages are available
- Might need to split the contig ([http://bioinfo.nhri.org.tw/cgi-bin/emboss/splitter](http://bioinfo.nhri.org.tw/cgi-bin/emboss/splitter))
- Limit searches to *Caudoviricetes* (taxid: 2731619) or the reference genome

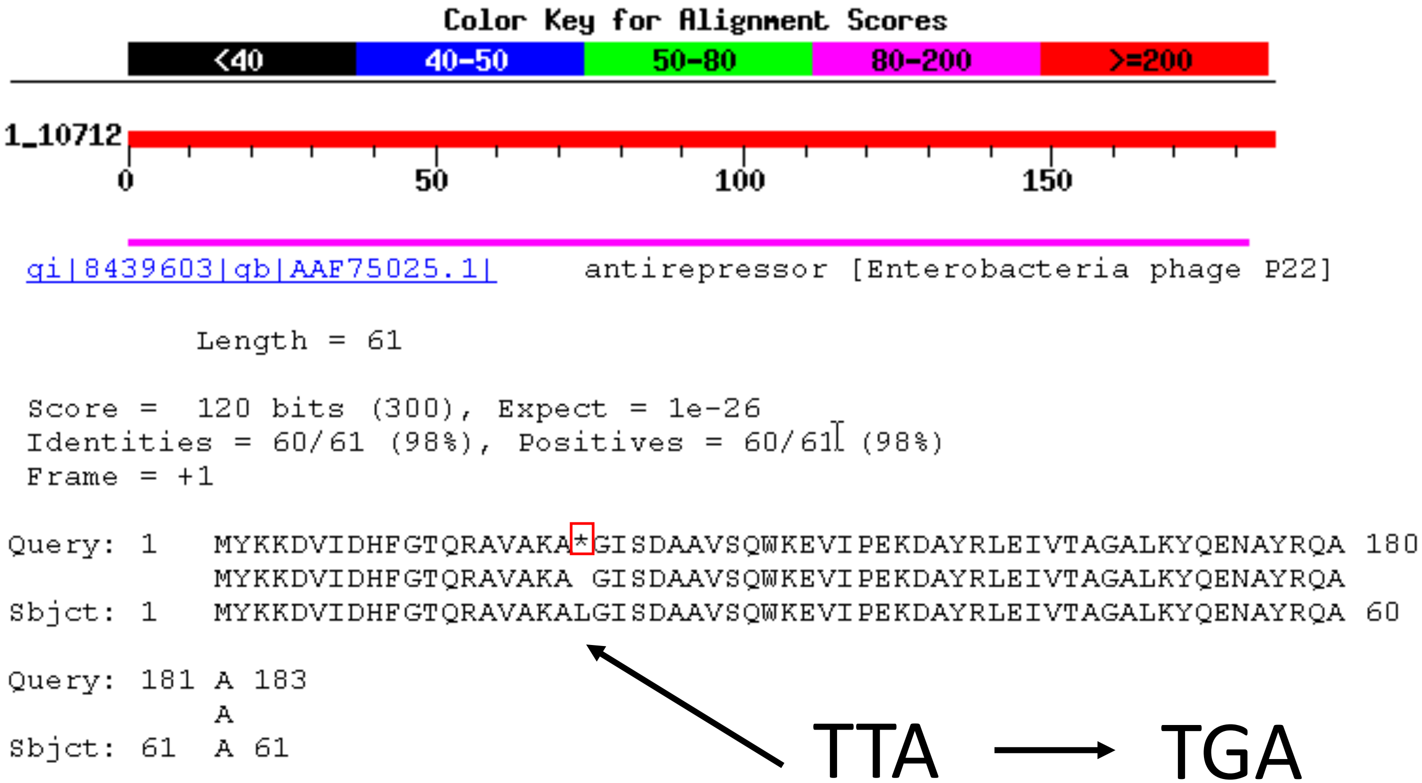


Distribution of 2 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_1993
0
50
100
150
200

```
□ gi|11611120|emb|CAC18561.1|    putative 0.45 protein [Bacteriophage phiYe03-12]

          Length = 66

 Score = 75.1 bits (183), Expect(2) = 4e-29
 Identities = 35/36 (97%), Positives = 36/36 (100%)
 Frame = +1

Query: 1   MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYGQ 108
            MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYG+
Sbjct: 1   MSKLLATSKIEGQCTVTLREYYHGSMGSTYVVRYGK 36

 Score = 74.3 bits (181), Expect(2) = 4e-29
 Identities = 31/31 (100%), Positives = 31/31 (100%)
 Frame = +2

Query: 107 KQVTHWVNPILAQEDYQSCVLHQTTCAGWND 199
            KQVTHWVNPILAQEDYQSCVLHQTTCAGWND
Sbjct: 36  KQVTHWVNPILAQEDYQSCVLHQTTCAGWND 66
```

# Internal Stop Codons

- Easy to miss using BLASTx
- Mis-called base substitutions can cause internal stop codons

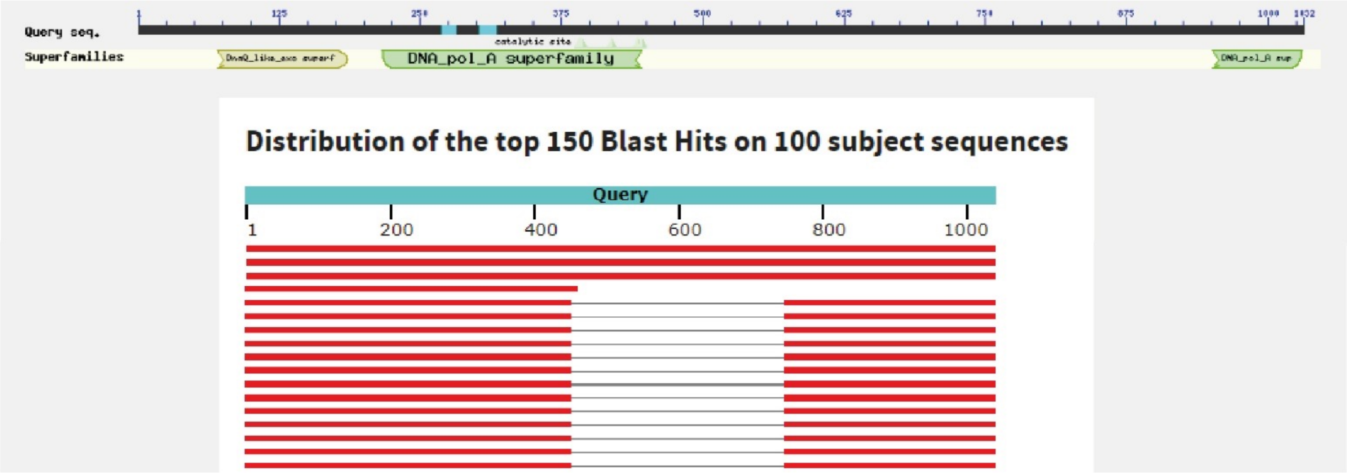


```
gi|8439603|gb|AAF75025.1|     antirepressor [Enterobacteria phage P22]

         Length = 61

 Score =  120 bits (300), Expect = 1e-26
 Identities = 60/61 (98%), Positives = 60/61 (98%)
 Frame = +1

Query: 1   MYKKDVIDHFGTQRAVAKA*GISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA 180
            MYKKDVIDHFGTQRAVAKA GISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA
Sbjct: 1   MYKKDVIDHFGTQRAVAKALGISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQA 60

Query: 181 A 183
            A
Sbjct: 61  A 61
```

TTA → TGA

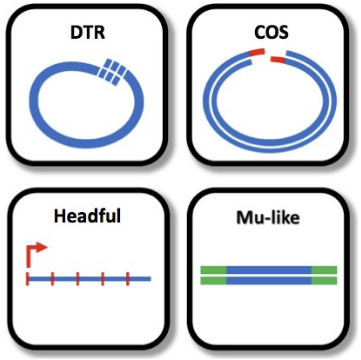
# Introns and Inteins

- Relatively rare
- Gene encoding the DNA polymerase in vB_SenS-Ent1. Some members of the *Jerseyvirinae* lack the intein coding region.



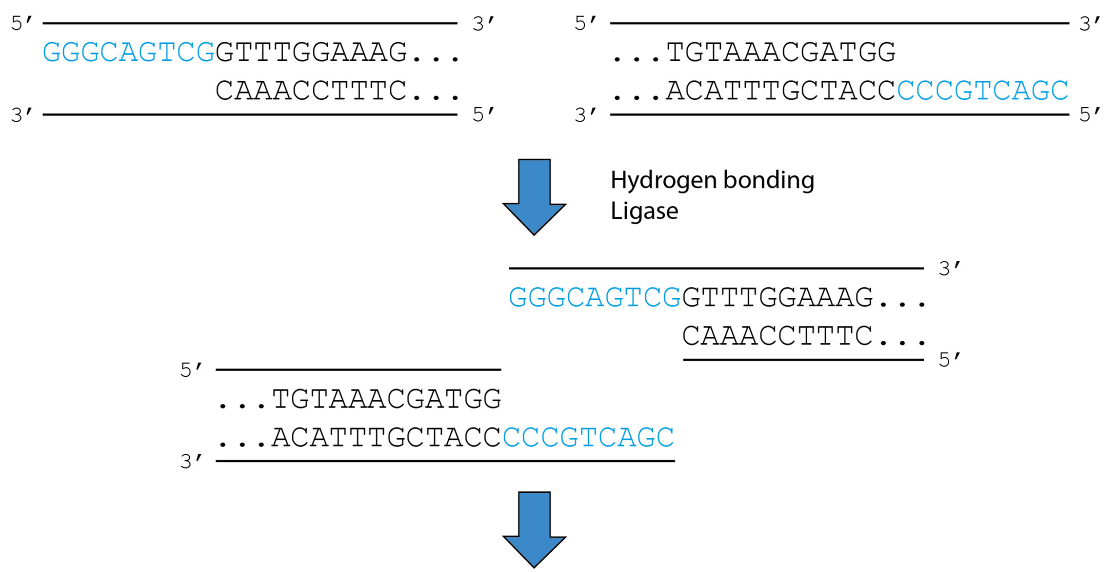
- Difficult to predict splice sites
  - InBase: https://inbase.ligsciss.com/iwai/InBase/tools.neb.com/inbase/identify.html
  - ISSPred: https://webs.iiitd.edu.in/raghava/isspred/index.html

# Genome Termini



- **Cohesive Ends – 5’ or 3’ extensions**
- Site specific packaging
- Determine by primer walking, annealing of restriction fragments (Casjens & Gilcrease, 2009; https://phagesdb.org/blog/posts/25/)

**5’ cohesive ends**



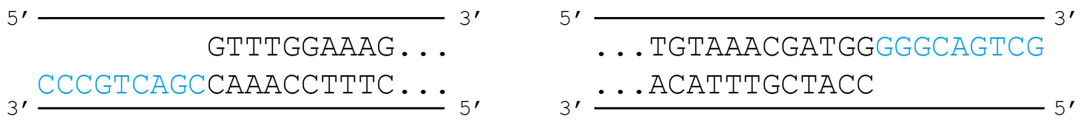
```
5′ ________________________ 3′    5′ ____________________________ 3′
GGGCAGTCGGTTTGGAAAG...            ...TGTAAACGATGG
         CAAACCTTTC...            ...ACATTTGCTACCCCCGTCAGC
3′ ________________________ 5′    3′ ____________________________ 5′
```

Hydrogen bonding
Ligase

```
                         ________________________ 3′
                         GGGCAGTCGGTTTGGAAAG...
                                  CAAACCTTTC...
5′ _______________                ______________ 5′
...TGTAAACGATGG
...ACATTTGCTACCCCCGTCAGC
3′ _________________________
```

Circular molecule

**3’ cohesive ends**

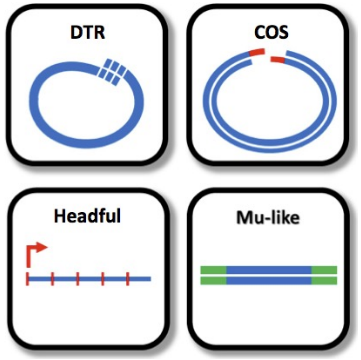


```
5′ ________________________ 3′    5′ ____________________________ 3′
         GTTTGGAAAG...            ...TGTAAACGATGGGGGCAGTCG
CCCGTCAGCCAAACCTTTC...            ...ACATTTGCTACC
3′ ________________________ 5′    3′ ____________________________ 5′
```

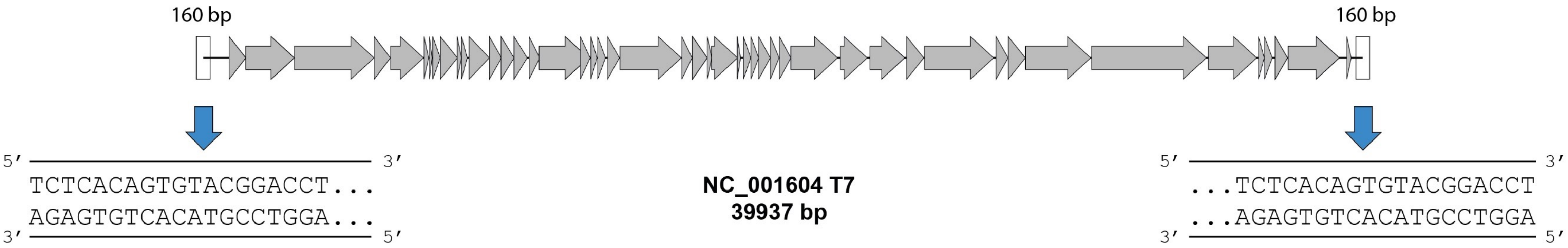
E.g. P2, Lambda

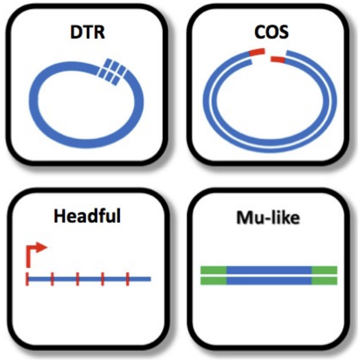
E.g. HK97, D3, many mycobacteriophages

# Genome Termini



- **Terminal redundancy – Direct repeats**
- *Autographiviridae* (e.g. T7, SP6, ϕKMV), T5, A511
- Vary in length (long/short repeats)
  - *Escherichia* phage T7 – 160 bp
  - Listeria phage A511 – 3,125 bp
  - *Escherichia phage T5 – 10,219 bp*
  - *Bacillus phage SPO1 – 13,185 bp*




DTR

COS

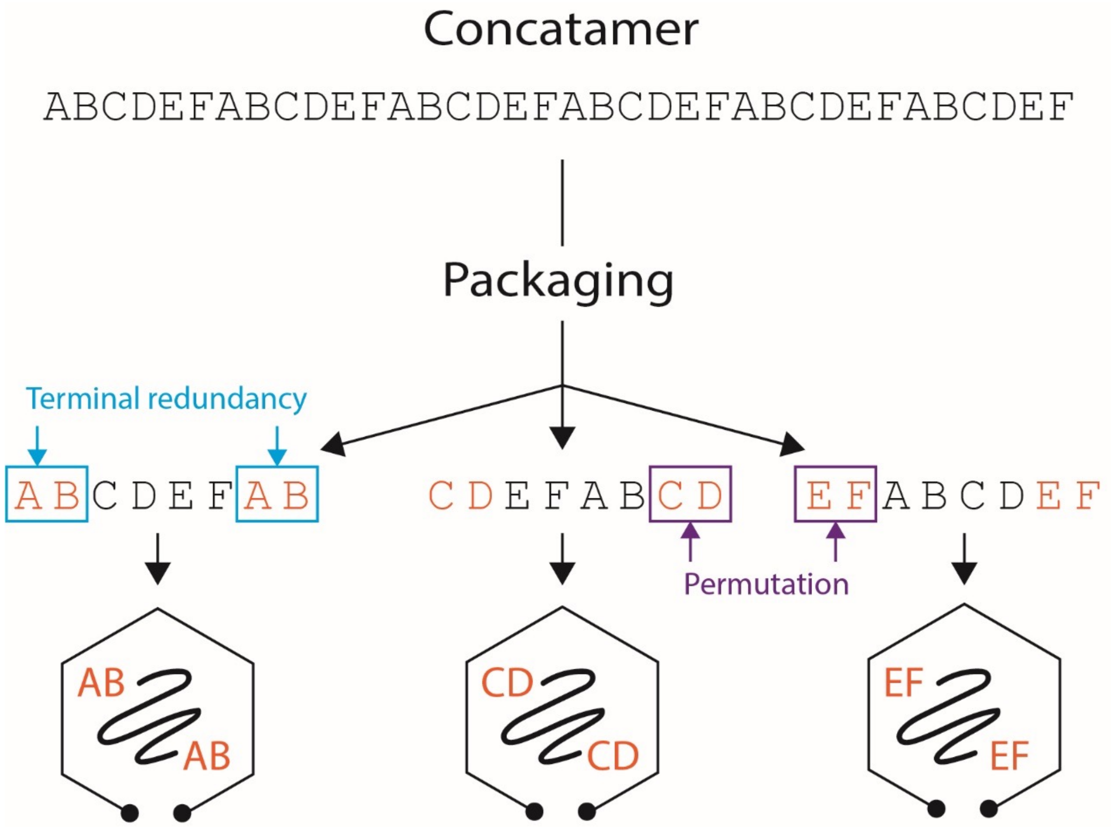
Headful

Mu-like

# Genome Termini

- **Terminal redundancy with circular permutation**
- T4, P1
- Characteristic of headful packaging
- Length of redundancy varies according to the phage

- Open genome according to convention
  - 1st nucleotide of small Terminase subunit
  - 1st nucleotide of rIIA


Concatamer

ABCDEFABCDEFABCDEFABCDEFABCDEFABCDEF

Packaging

Terminal redundancy

ABCDEFAB

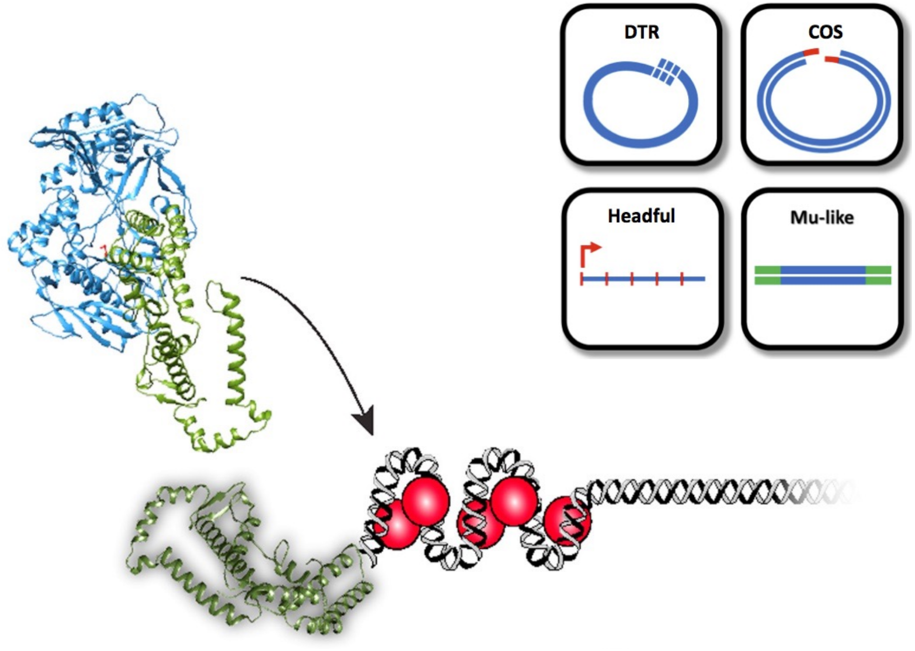
CDEFABCD

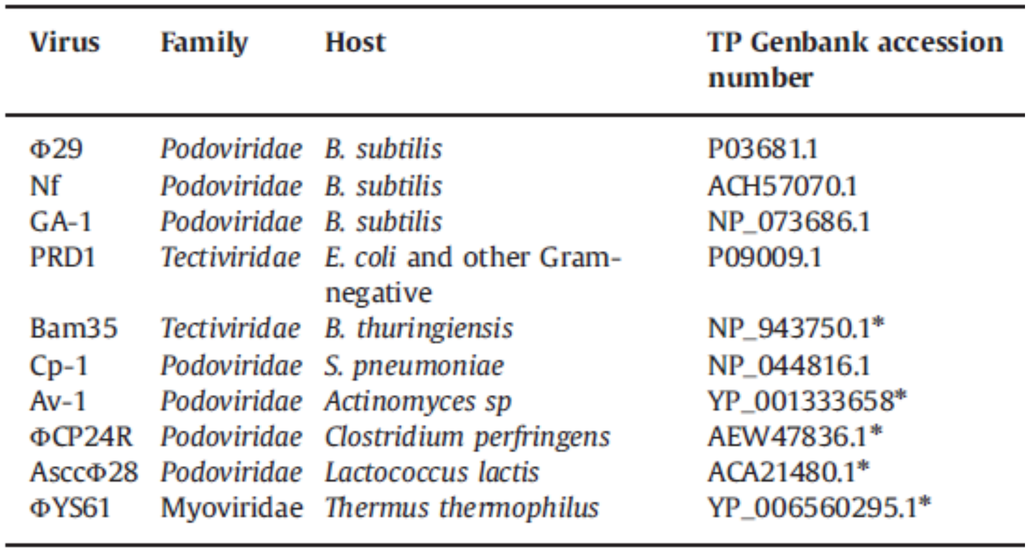
EFABCDEF

Permutation

# Genome Termini

- **Terminal proteins**
- Protein-primed replication of linear dsDNA
- Terminal proteins show low sequence homology
- Requires in vitro approaches
  - Migration in gels +/- protease treatment

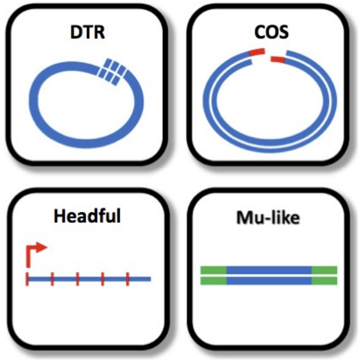




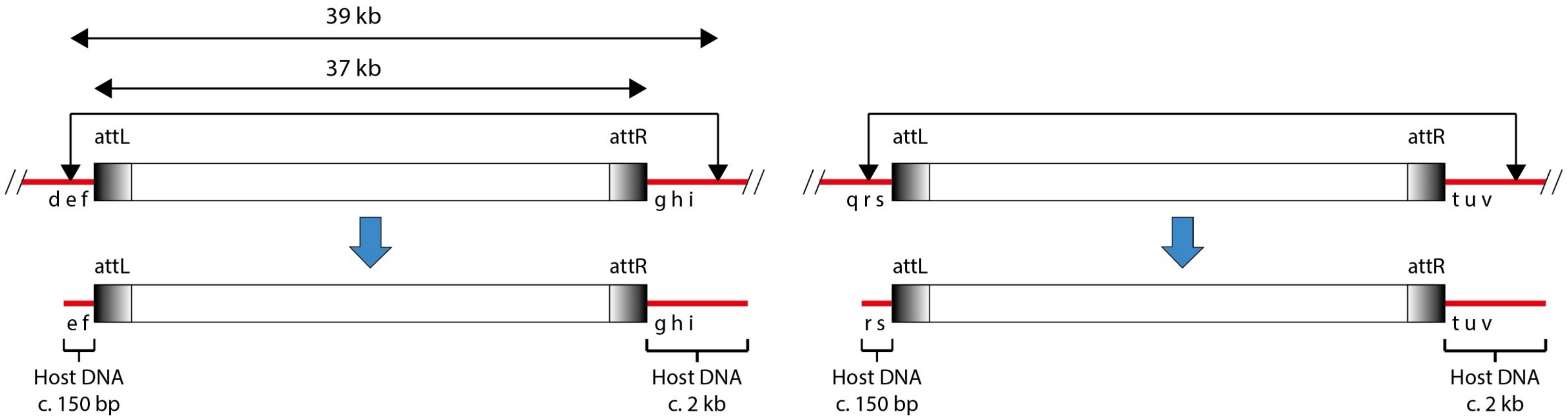
| Virus | Family | Host | TP Genbank accession number |
|---|---|---|---|
| Φ29 | *Podoviridae* | *B. subtilis* | P03681.1 |
| Nf | *Podoviridae* | *B. subtilis* | ACH57070.1 |
| GA-1 | *Podoviridae* | *B. subtilis* | NP_073686.1 |
| PRD1 | *Tectiviridae* | *E. coli* and other Gram-negative | P09009.1 |
| Bam35 | *Tectiviridae* | *B. thuringiensis* | NP_943750.1* |
| Cp-1 | *Podoviridae* | *S. pneumoniae* | NP_044816.1 |
| Av-1 | *Podoviridae* | *Actinomyces sp* | YP_001333658* |
| ΦCP24R | *Podoviridae* | *Clostridium perfringens* | AEW47836.1* |
| AsccΦ28 | *Podoviridae* | *Lactococcus lactis* | ACA21480.1* |
| ΦYS61 | Myoviridae | *Thermus thermophilus* | YP_006560295.1* |

Virology (2014), 468-470:322

# Genome Termini

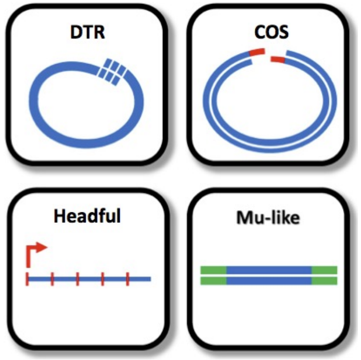


- **Host DNA**
- Replicative transposition – Mu, D108, B3 and others
- Random integration results in variable ends of host DNA
- B3/Mu: primer walk-out strategy – loss of base identification after terminal 5’-TG dinucleotides



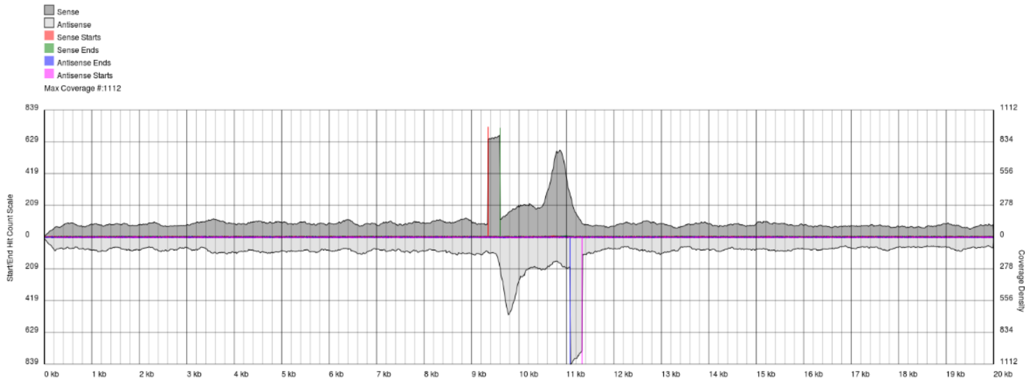
Adapted from http://www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/transposons/Mu.html

# Computational Prediction of Termini

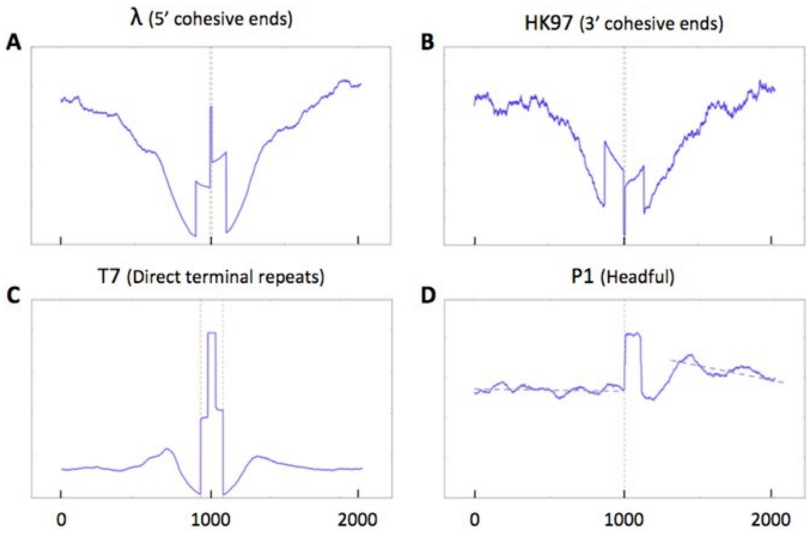


- Use biases in numbers of reads

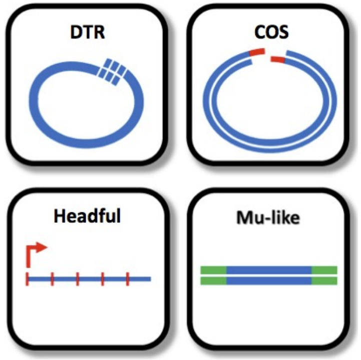
- PAUSE (Pileup Analysis Using Starts and Ends)
  - Center for Phage Technology



- PhageTerm
  - Requires assembled genome and sequence reads



# Genome Termini: Lab methods



- Restriction sites
  - NEBcutter (https://nc3.neb.com/NEBcutter/)
  - Do the predicted fragments from the assembly exist physically?

- BAL-31 exonuclease
  - Fragments with defined ends will show a reduction in length
  - Circularly permuted ends will show

- Fast/slow cooling
  - Annealing of fragments with cohesive ends – can be problematic depending upon sequence composition
- Sanger sequencing
  - Walk-out methods from genome termini

# The final(ish) product

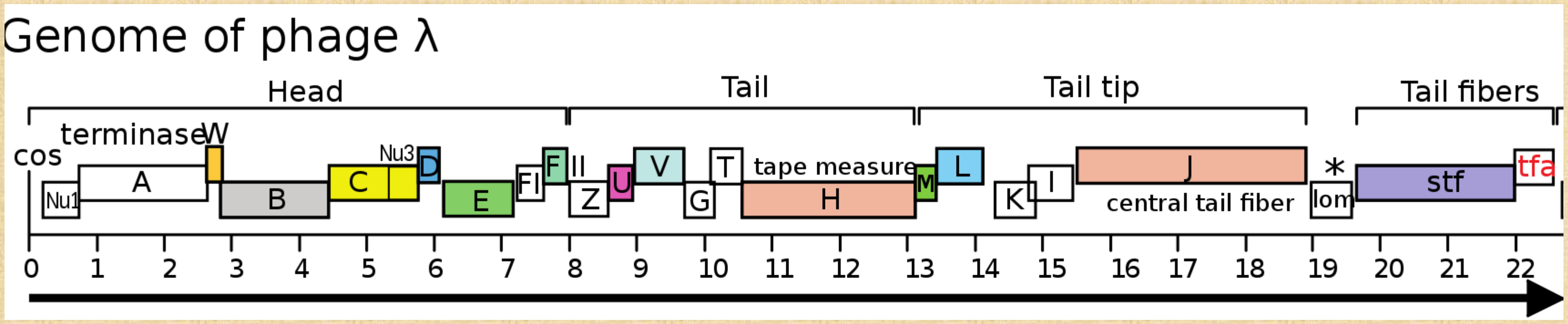
- I have a finalised genome, what’s next?

- Annotation (Andrew Kropinski)
  - What genes does my phage code for?
  - What are the gene products?

- Classification (Evelien Adriaenssens)
  - Where does my phage fall in the phage biosphere?
  - Is it new or is it a representative of an existing family/genus/species?

# Resources

- http://phagesdb.org/workflow/Sequencing/
- Shen & Millard (2021) PHAGE, 2(4):183
- http://millardlab.org/lab-members/alumni/lucy-gannon/lucys-beginner-guide-to-bacteriophage-genome-assembly/
- Russell (2018) Methods in Molecular Biology, 1681:109
- Turner, Adriaenssens, Tolstoy, Kropinski (2021) PHAGE, 2(4)170
- Online Analysis Tools: http://molbiol-tools.ca (thank you Andrew!)
- CPT Phage Galaxy: https://cpt.tamu.edu/galaxy-pub
- CLIMB: https://www.climb.ac.uk/getting-started/



# Genes in Phage Genomes

Andrew M. Kropinski

Phage.Canada@gmail.com

UNIVERSITY of GUELPH

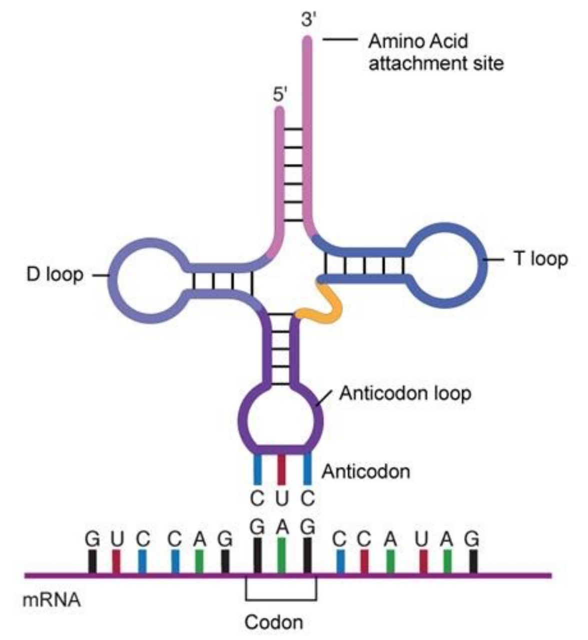
# Genes

❑Identification of tRNA-encoding sequences

❑Identification of open reading frames (ORFs) coding for proteins (CDSs)

**N.B.** CDSs and tRNA genes don’t overlap

# tRNAs in Phage Genomes



- ❑ Can be found using:
  - tRNAscan-SE 2.0 (http://lowelab.ucsc.edu/tRNAscan-SE/)
  - ARAGORN (http://130.235.46.10/ARAGORN/)
- ❑ Please note that occasionally automated annotation programs miss tRNAs (e.g. MyRAST)

# ORF vs CDS

❑ an ORF is a sequence that has a length divisible by three and is bounded by stop codons

❑ stop codons - TAA, TAG or TGA

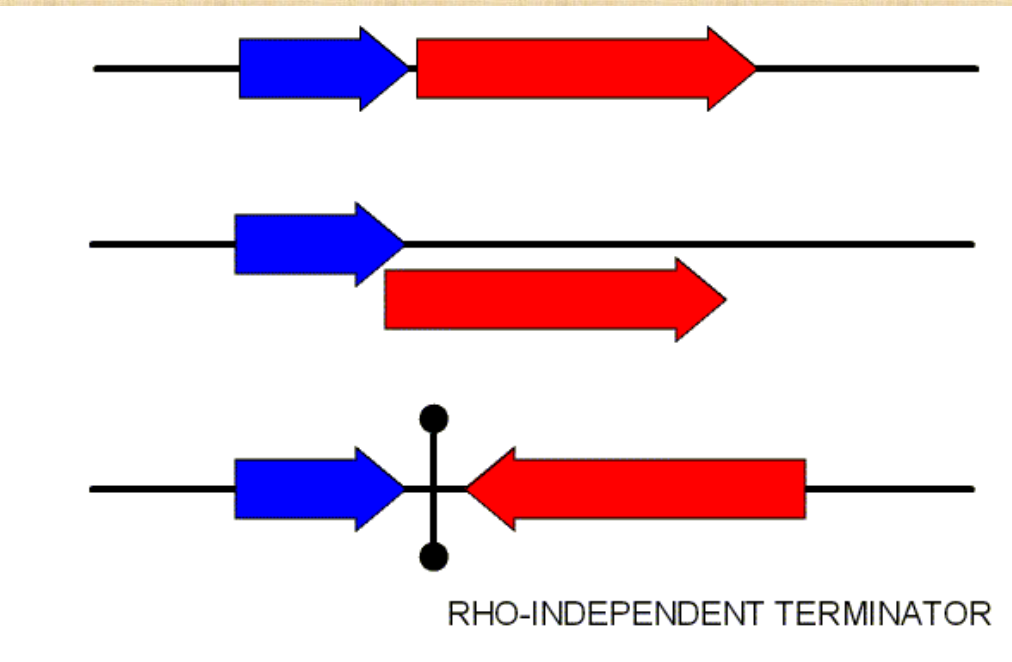
❑ may not specify a protein

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

❑ CDS has an important upstream feature – ribosome-binding site or Shine-Dalgarno box (GGAGGT)

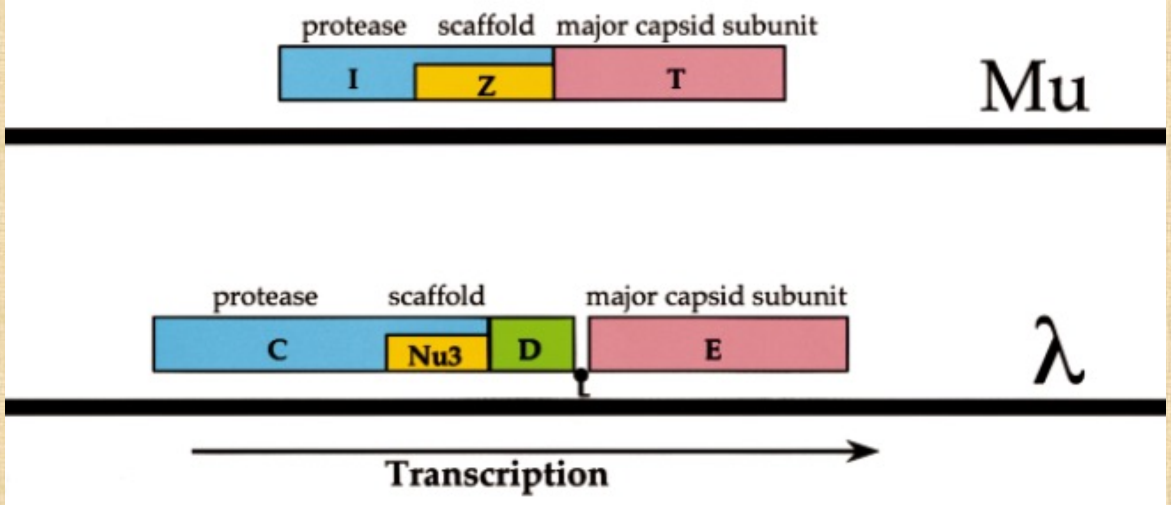
# Arrangement of Genes

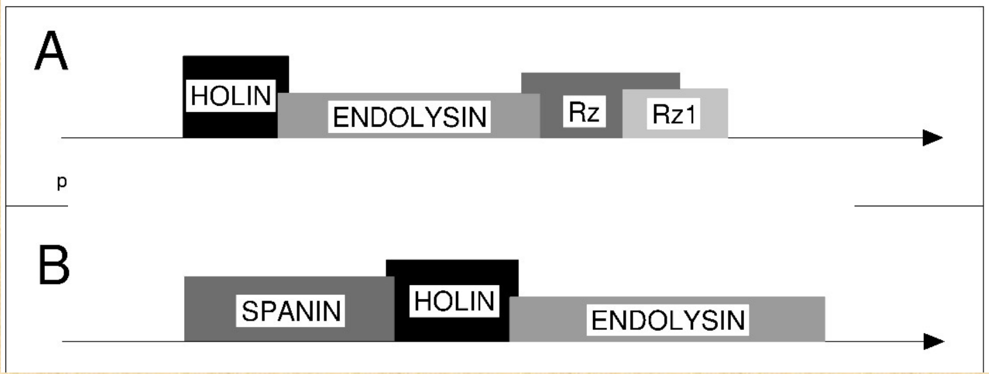
- Common



# Arrangement of Genes (cont.)

- Rare – heavily overlapped or embedded genes





- More common in the case of the lysis cassette

# Automated Annotation

- A good way to start
- Web:
  - RAST (http://rast.nmpdr.org/)*
  - DFAST (https://dfast.nig.ac.jp/)
  - PATRIC (https://www.patricbrc.org/app/Annotation)* – uses RASTtk
  - PROKKA* (https://kbase.us/applist/apps/ProkkaAnnotation/annotate_contigs/release?gclid=EAIaIQobChMI-93RvvOJ-AIVGxXUAR2e4gTBEAAYASAAEgJWw_D_BwE)

  * requires free registration
- DFAST is incredibly fast, the others depend upon how busy the server is.
- desired output – GenBank flatfile (*.gb or *.gbk)

# Comments on Autoannotation

❑ Can you believe the autoannotation results?

No:

a) Adequate at defining correct initiation codons

b) Adequate at defining product function

c) But, bad at identifying small CDSs

- Insertion of missed genes – e.g. λ Ral (28 aa) and Sf6 gp45 (27 aa)
- Correction for wrong initiation codons

  RBS INITIATION CODON

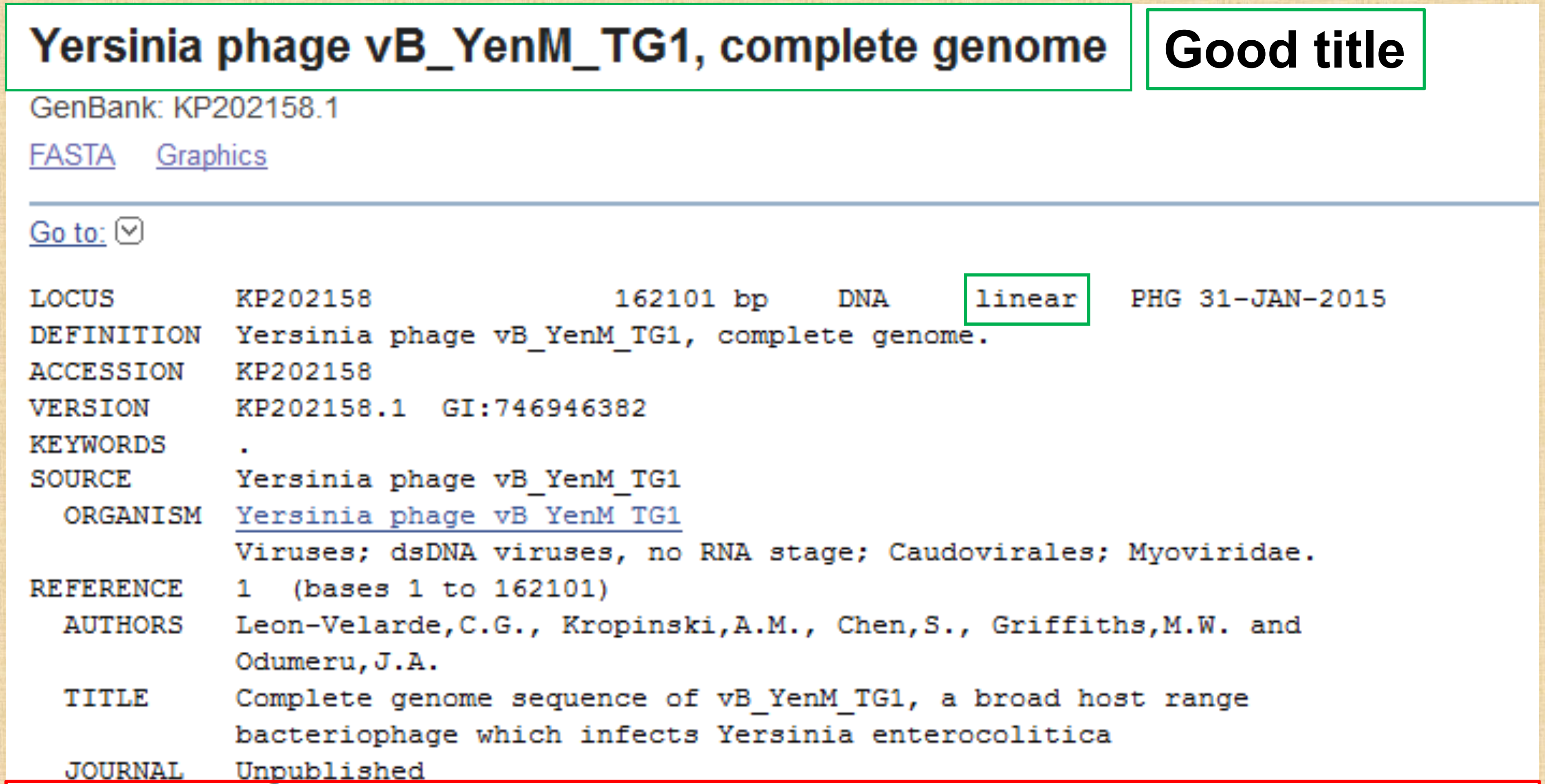
  GGAGGT (N3-10) ATG(GTG,TTG)xxxx
- Correction of names of annotated genes products

# Freeware for Manual Genomic Annotation

- ❑ Artemis – old and reliable (Unix, PC, Mac)
  http://www.sanger.ac.uk/science/tools/artemis
- ❑ DNA Master – used by the SEA PHAGES group
  https://seaphages.org/blog/2016/11/16/dna-master-updated-use-secure-ncbi-connections/
- ❑ UGENE – continually updated (Unix, PC, Mac)
  http://ugene.net/

  - What you want minimally is software which will display DNA sequence and the translated sequence (protein) simultaneously

# Accurate GenBank File



Yersinia phage vB_YenM_TG1, complete genome

Good title

GenBank: KP202158.1

FASTA Graphics

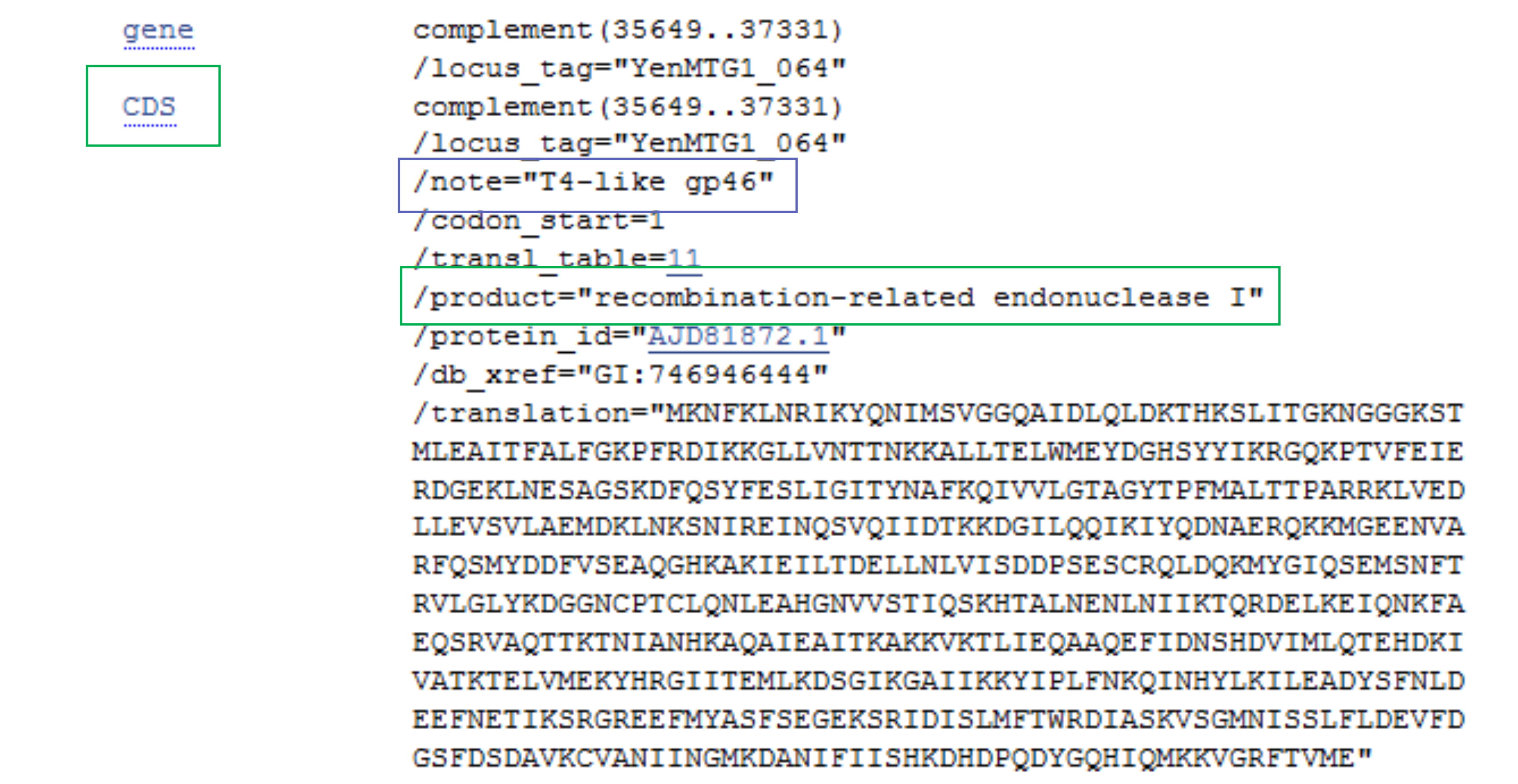
Go to:

```
LOCUS       KP202158              162101 bp    DNA     linear   PHG 31-JAN-2015
DEFINITION  Yersinia phage vB_YenM_TG1, complete genome.
ACCESSION   KP202158
VERSION     KP202158.1  GI:746946382
KEYWORDS    .
SOURCE      Yersinia phage vB_YenM_TG1
  ORGANISM  Yersinia phage vB_YenM_TG1
            Viruses; dsDNA viruses, no RNA stage; Caudovirales; Myoviridae.
REFERENCE   1  (bases 1 to 162101)
  AUTHORS   Leon-Velarde,C.G., Kropinski,A.M., Chen,S., Griffiths,M.W. and
            Odumeru,J.A.
  TITLE     Complete genome sequence of vB_YenM_TG1, a broad host range
            bacteriophage which infects Yersinia enterocolitica
  JOURNAL   Unpublished
```

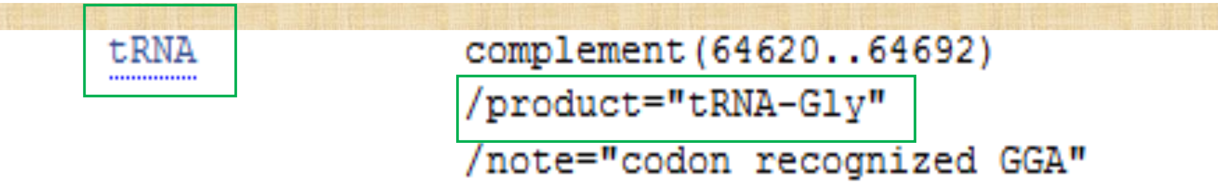
- Bacteriophage LKD16 complete genome, specific host Pseudomonas aeruginosa
- Pseudomonas phage phi-2, complete genome, isolated from Pseudomonas fluorescens SBW25

Circular

# Accurate GenBank File 2

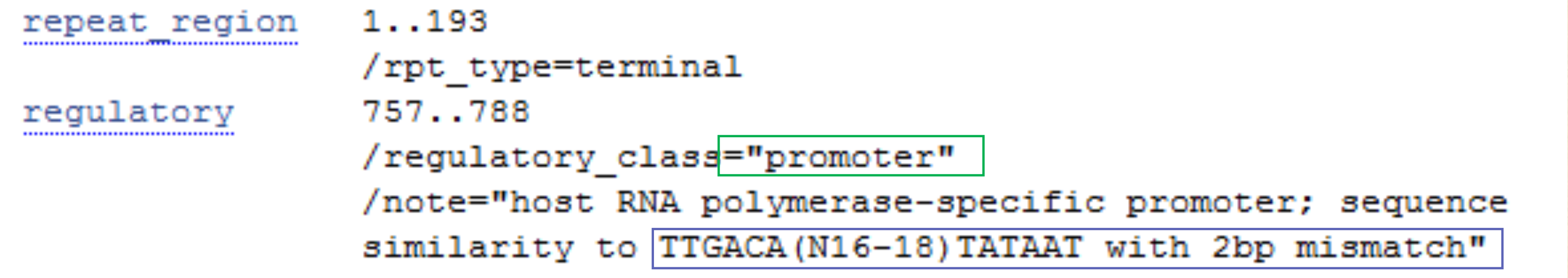


```
     gene            complement(35649..37331)
                     /locus_tag="YenMTG1_064"
     CDS             complement(35649..37331)
                     /locus_tag="YenMTG1_064"
                     /note="T4-like gp46"
                     /codon_start=1
                     /transl_table=11
                     /product="recombination-related endonuclease I"
                     /protein_id="AJD81872.1"
                     /db_xref="GI:746946444"
                     /translation="MKNFKLNRIKYQNIMSVGGQAIDLQLDKTHKSLITGKNGGGKST
                     MLEAITFALFGKPFRDIKKGLLVNTTNKKALLTELWMEYDGHSYYIKRGQKPTVFEIE
                     RDGEKLNESAGSKDFQSYFESLIGITYNAFKQIVVLGTAGYTPFMALTTPARRKLVED
                     LLEVSVLAEMDKLNKSNIREINQSVQIIDTKKDGILQQIKIYQDNAERQKKMGEENVA
                     RFQSMYDDFVSEAQGHKAKIEILTDELLNLVISDDPSESCRQLDQKMYGIQSEMSNFT
                     RVLGLYKDGGNCPTCLQNLEAHGNVVSTIQSKHTALNENLNIIKTQRDELKEIQNKFA
                     EQSRVAQTTKTNIANHKAQAIEAITKAKKVKTLIEQAAQEFIDNSHDVIMLQTEHDKI
                     VATKTELVMEKYHRGIITEMLKDSGIKGAIIKKYIPLFNKQINHYLKILEADYSFNLD
                     EEFNETIKSRGREEFMYASFSEGEKSRIDISLMFTWRDIASKVSGMNISSLFLDEVFD
                     GSFDSDAVKCVANIINGMKDANIFIISHKDHDPQDYGQHIQMKKVGRFTVME"
```

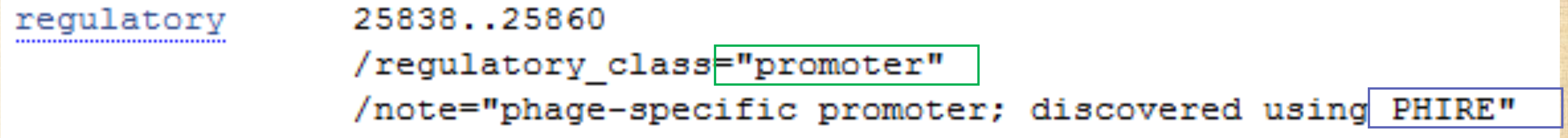


```
     tRNA            complement(64620..64692)
                     /product="tRNA-Gly"
                     /note="codon recognized GGA"
```

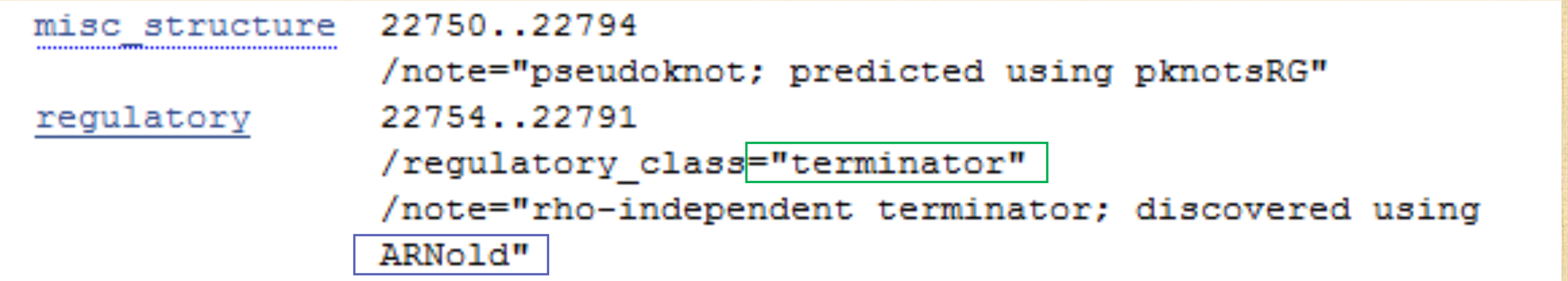
# Accurate GenBank File 3



```
repeat_region   1..193
                /rpt_type=terminal
regulatory      757..788
                /regulatory_class="promoter"
                /note="host RNA polymerase-specific promoter; sequence
                similarity to TTGACA(N16-18)TATAAT with 2bp mismatch"
```



```
regulatory      25838..25860
                /regulatory_class="promoter"
                /note="phage-specific promoter; discovered using PHIRE"
```



```
misc_structure  22750..22794
                /note="pseudoknot; predicted using pknotsRG"
regulatory      22754..22791
                /regulatory_class="terminator"
                /note="rho-independent terminator; discovered using
                ARNold"
```

# Locus tag

❑The locus_tag is a systematic gene identifier that is assigned to each gene. Each genome project have the same unique locus_tag prefix to ensure that a locus_tag is specific for a particular genome project. The locus_tag prefix must be 3-12 alphanumeric characters and the first character may not be a digit. Additionally locus_tag prefixes are case-sensitive. The locus_tag prefix is followed by an underscore and then an alphanumeric identification number that is unique within the given genome. Other than the single underscore used to separate the prefix from the identification number, no other special characters can be used in the locus_tag. Locus_tags must only be used in combination with a gene feature.

(https://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf)

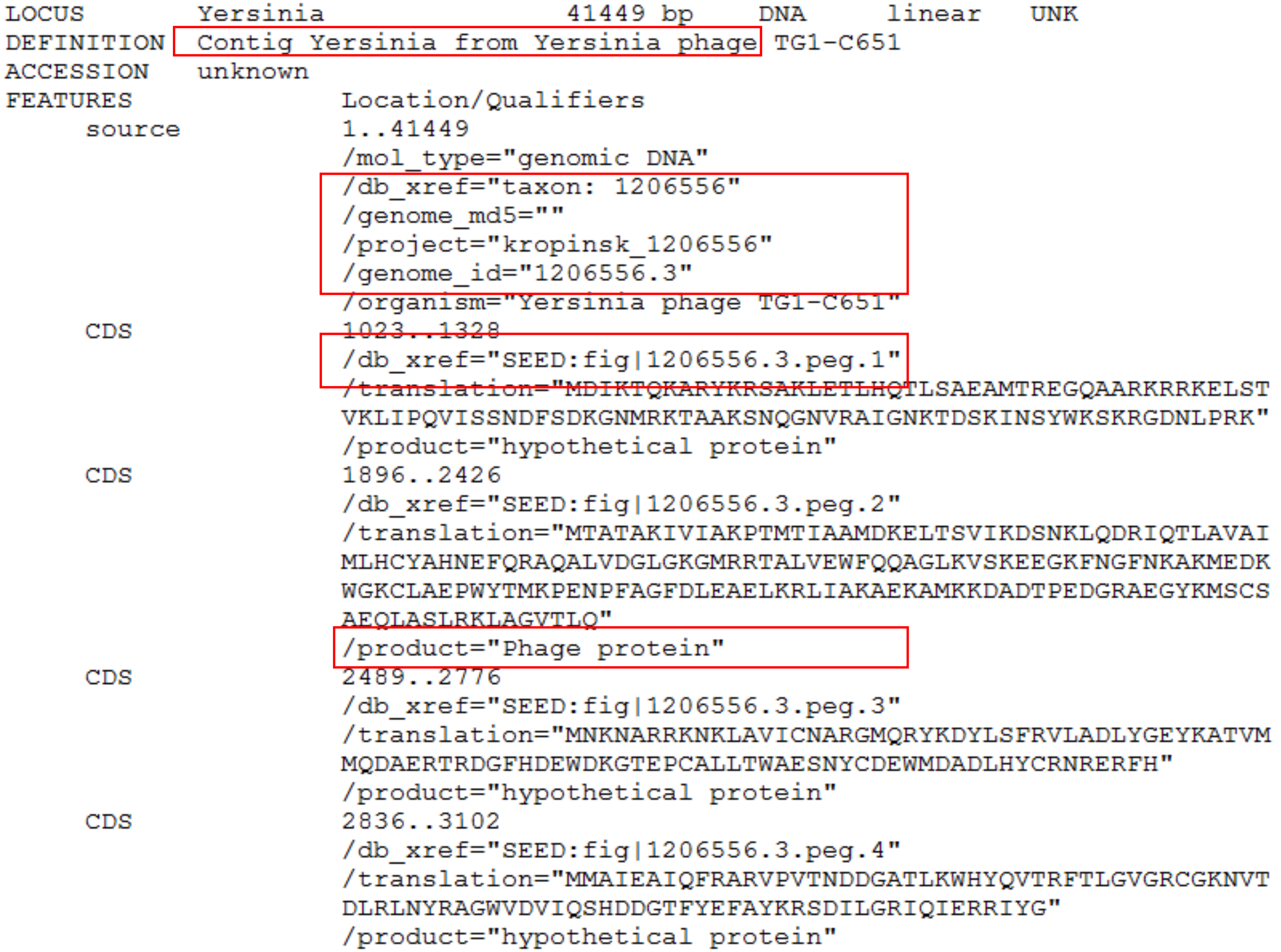
❑Use you phage name as the locus tag.

❑Not added by RAST, DFAST or PATRIC

# Massaging *.gbk files

- ❑ You will have to do this in all cases
- ❑ Be suspicious of gaps
- ❑ are protein homologs the same size
- ❑ do you have homing  endonucleases – be suspicious of fragmented genes

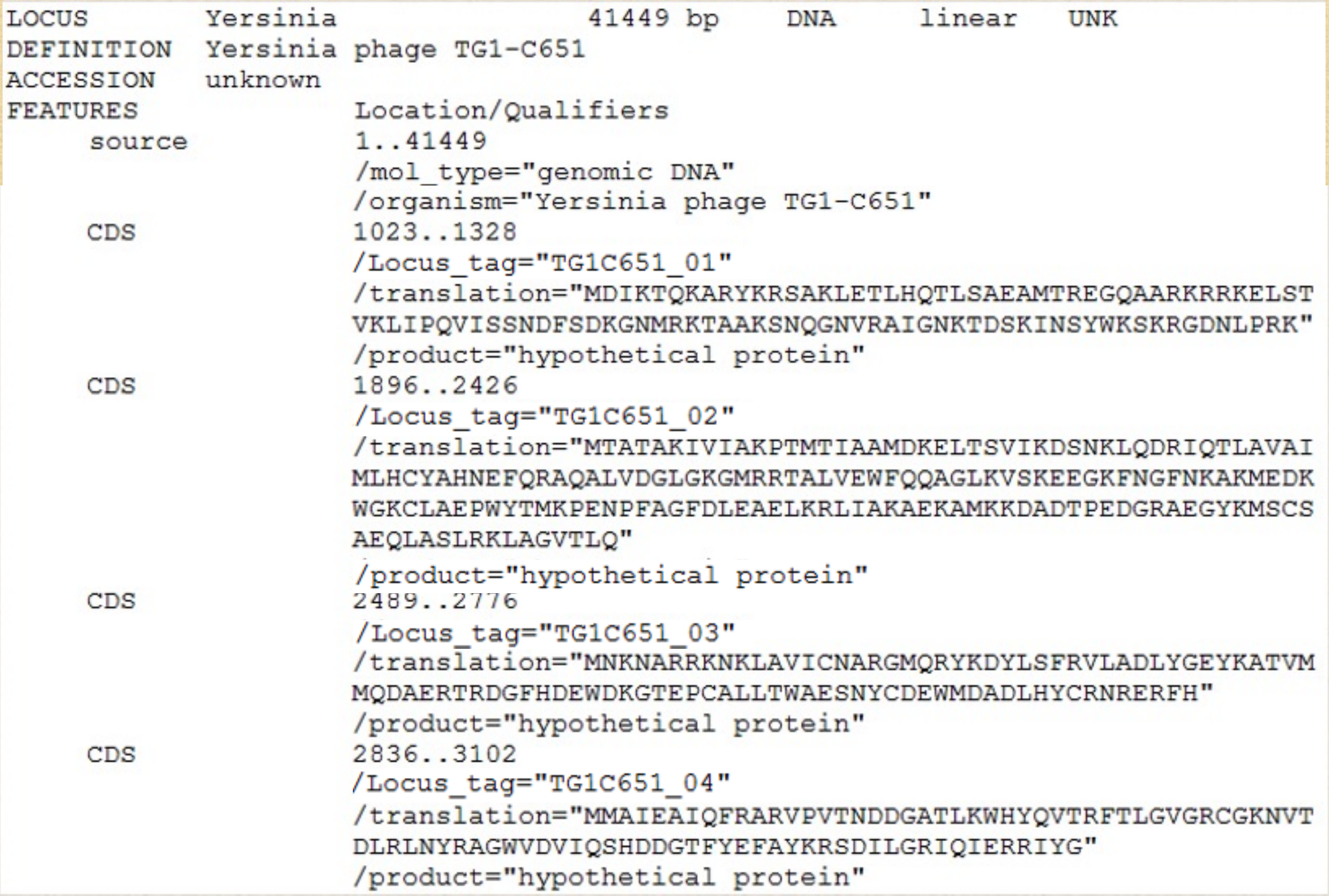
# Massaging RAST Data



```
LOCUS       Yersinia                 41449 bp    DNA     linear   UNK
DEFINITION  Contig Yersinia from Yersinia phage TG1-C651
ACCESSION   unknown
FEATURES             Location/Qualifiers
     source          1..41449
                     /mol_type="genomic DNA"
                     /db_xref="taxon: 1206556"
                     /genome_md5=""
                     /project="kropinsk_1206556"
                     /genome_id="1206556.3"
                     /organism="Yersinia phage TG1-C651"
     CDS             1023..1328
                     /db_xref="SEED:fig|1206556.3.peg.1"
                     /translation="MDIKTQKARYKRSAKLETLHQTLSAEAMTREGQAARKRRKELST
                     VKLIPQVISSNDFSDKGNMRKTAAKSNQGNVRAIGNKTDSKINSYWKSKRGDNLPRK"
                     /product="hypothetical protein"
     CDS             1896..2426
                     /db_xref="SEED:fig|1206556.3.peg.2"
                     /translation="MTATAKIVIAKPTMTIAAMDKELTSVIKDSNKLQDRIQTLAVAI
                     MLHCYAHNEFQRAQALVDGLGKGMRRTALVEWFQQAGLKVSKEEGKFNGFNKAKMEDK
                     WGKCLAEPWYTMKPENPFAGFDLEAELKRLIAKAEKAMKKDADTPEDGRAEGYKMSCS
                     AEQLASLRKLAGVTLQ"
                     /product="Phage protein"
     CDS             2489..2776
                     /db_xref="SEED:fig|1206556.3.peg.3"
                     /translation="MNKNARRKNKLAVICNARGMQRYKDYLSFRVLADLYGEYKATVM
                     MQDAERTRDGFHDEWDKGTEPCALLTWAESNYCDEWMDADLHYCRNRERFH"
                     /product="hypothetical protein"
     CDS             2836..3102
                     /db_xref="SEED:fig|1206556.3.peg.4"
                     /translation="MMAIEAIQFRARVPVTNDDGATLKWHYQVTRFTLGVGRCGKNVT
                     DLRLNYRAGWVDVIQSHDDGTFYEFAYKRSDILGRIQIERRIYG"
                     /product="hypothetical protein"
```

**Neat but definition wrong & no locus tags or gene identifiers in WordPad**

# Massaging RAST Data 3



```
LOCUS       Yersinia               41449 bp    DNA     linear   UNK
DEFINITION  Yersinia phage TG1-C651
ACCESSION   unknown
FEATURES             Location/Qualifiers
     source          1..41449
                     /mol_type="genomic DNA"
                     /organism="Yersinia phage TG1-C651"
     CDS             1023..1328
                     /Locus_tag="TG1C651_01"
                     /translation="MDIKTQKARYKRSAKLETLHQTLSAEAMTREGQAARKRRKELST
                     VKLIPQVISSNDFSDKGNMRKTAAKSNQGNVRAIGNKTDSKINSYWKSKRGDNLPRK"
                     /product="hypothetical protein"
     CDS             1896..2426
                     /Locus_tag="TG1C651_02"
                     /translation="MTATAKIVIAKPTMTIAAMDKELTSVIKDSNKLQDRIQTLAVAI
                     MLHCYAHNEFQRAQALVDGLGKGMRRTALVEWFQQAGLKVSKEEGKFNGFNKAKMEDK
                     WGKCLAEPWYTMKPENPFAGFDLEAELKRLIAKAEKAMKKDADTPEDGRAEGYKMSCS
                     AEQLASLRKLAGVTLQ"
                     /product="hypothetical protein"
     CDS             2489..2776
                     /Locus_tag="TG1C651_03"
                     /translation="MNKNARRKNKLAVICNARGMQRYKDYLSFRVLADLYGEYKATVM
                     MQDAERTRDGFHDEWDKGTEPCALLTWAESNYCDEWMDADLHYCRNRERFH"
                     /product="hypothetical protein"
     CDS             2836..3102
                     /Locus_tag="TG1C651_04"
                     /translation="MMAIEAIQFRARVPVTNDDGATLKWHYQVTRFTLGVGRCGKNVT
                     DLRLNYRAGWVDVIQSHDDGTFYEFAYKRSDILGRIQIERRIYG"
                     /product="hypothetical protein"
```

**Perfect**

# Comments on Autoannotation

❑ Can you believe the autoannotation results?

No:

a) Adequate at defining correct initiation codons

b) Adequate at defining product function

c) But, bad at identifying small CDSs

- Insertion of missed genes – e.g. λ Ral (28 aa) and Sf6 gp45 (27 aa)
- Correction for wrong initiation codons

  RBS INITIATION CODON

  GGAGGT (N3-10) ATG(GTG,TTG)xxxx
- Correction of names of annotated genes products

# Comments on Autoannotation 2

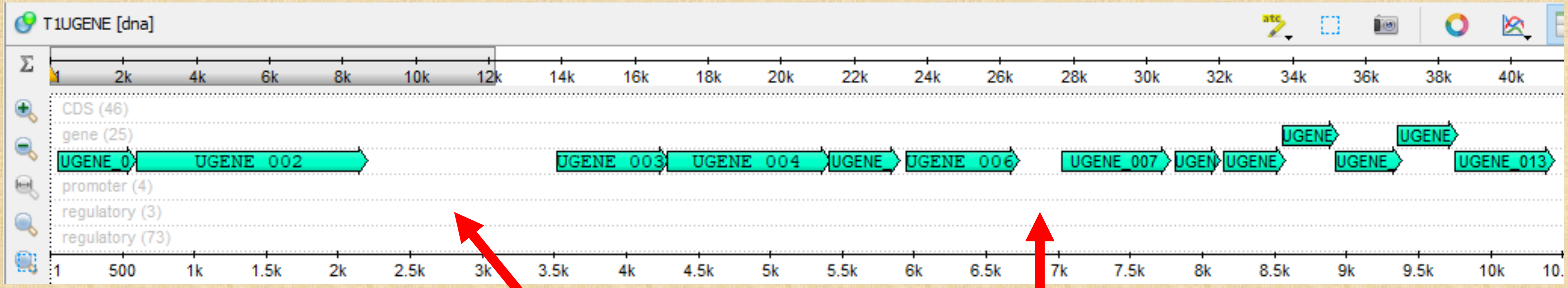
❑ What next?

“Manual” checking of results using software package that will present DNA sequence and overlay CDSs:

- Artemis: Genome Browser and Annotation Tool
- DNA Master
- Unipro UGENE (http://ugene.net/)

# Using UGENE to proof-read

❑Open *.gbk file in UGENE

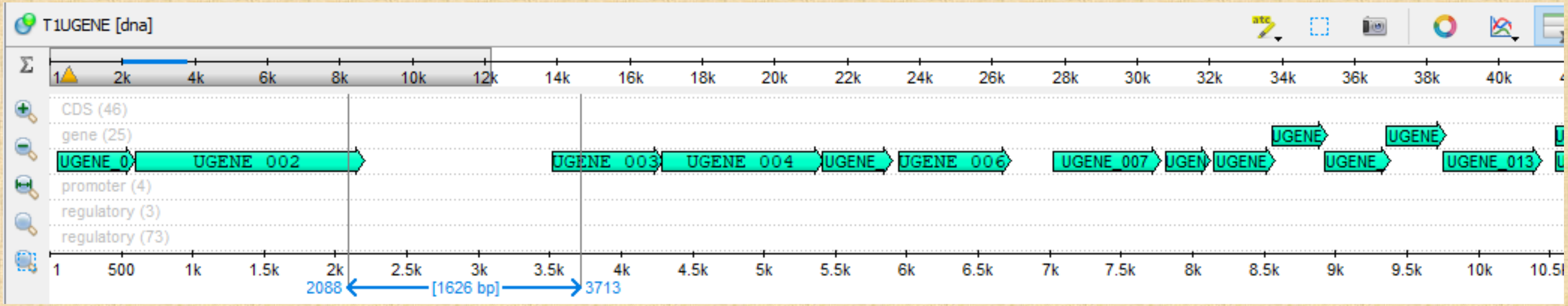


Gaps are interesting!
Is something missing?

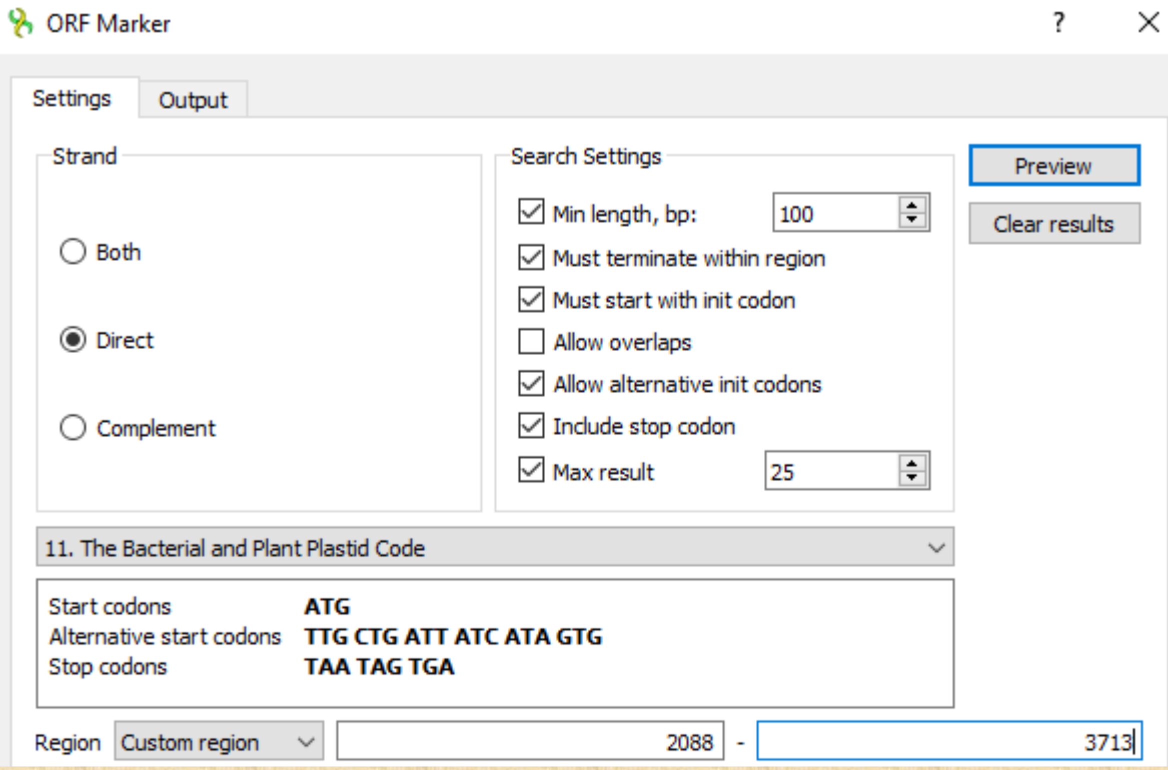
❑Two possibilities:

- Missing CDS
- Upstream initiation codon

# Using UGENE to proof-read 2



- ORF Marker



Strand

○ Both

● Direct

○ Complement

Search Settings

☑ Min length, bp: 100

☑ Must terminate within region

☑ Must start with init codon

□ Allow overlaps

☑ Allow alternative init codons

☑ Include stop codon

☑ Max result 25

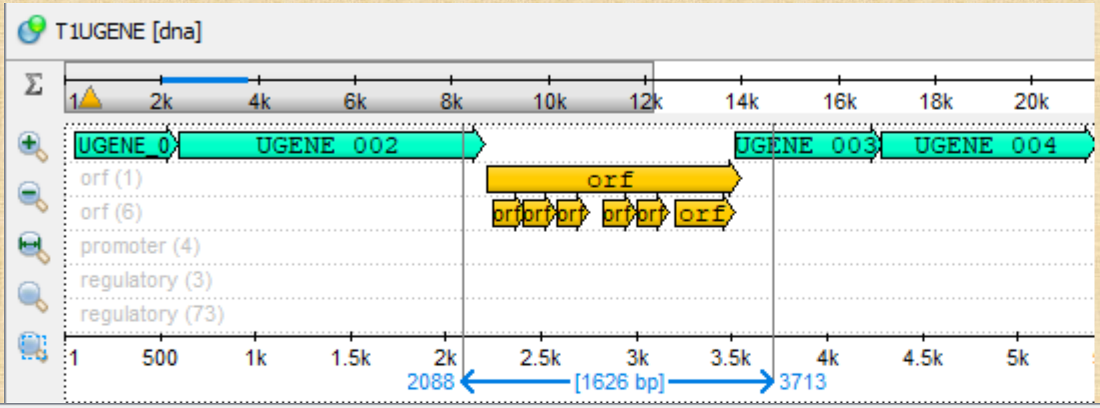
11. The Bacterial and Plant Plastid Code

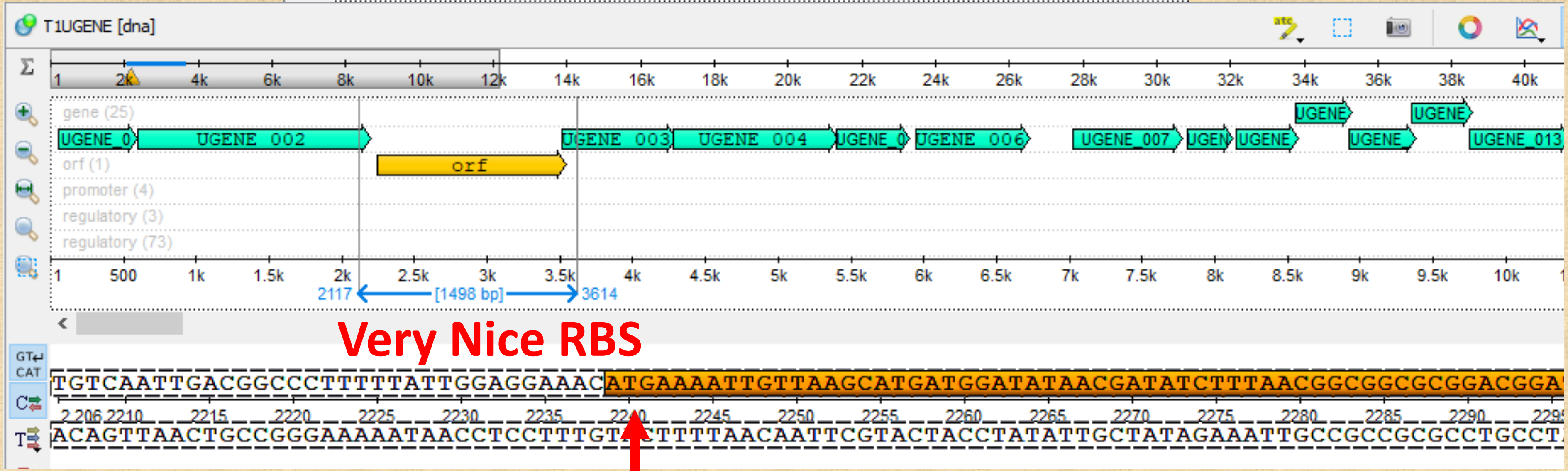
| | |
|---|---|
| Start codons | **ATG** |
| Alternative start codons | **TTG CTG ATT ATC ATA GTG** |
| Stop codons | **TAA TAG TGA** |

Region Custom region 2088 - 3713

# Using UGENE to proof-read 3

- ORF Marker





**Very Nice RBS**

```
TGTCAATTGACGGCCCTTTTTATTGGAGGAAACATGAAAATTGTTAAGCATGATGGATATAACGATATCTTTAACGGCGGCGCGGACGGA
 2206 2210  2215  2220  2225  2230  2235  2240  2245  2250  2255  2260  2265  2270  2275  2280  2285  2290  229
ACAGTTAACTGCCGGGAAAAATAACCTCCTTTGTACTTTTAACAATTCGTACTACCTATATTGCTATAGAAATTGCCGCCGCGCCTGCCT
```

# Section 2 – naming gene products

# What do I call the gene product (i.e. phage protein)?

❑“phage hypothetical protein” – redundant

❑“gp87” (gp = gene product) → hypothetical protein

- 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

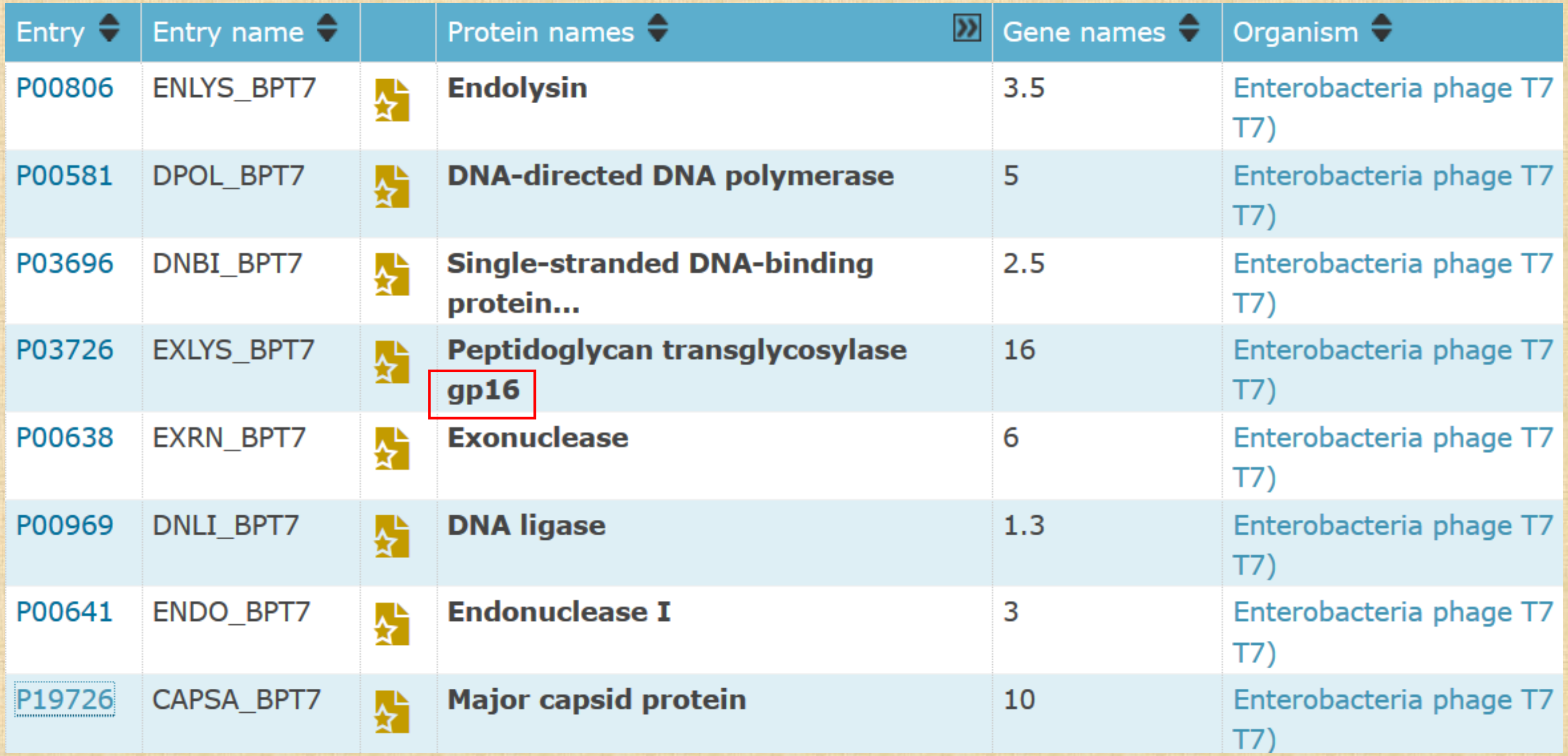
❑ /product="UboA"; "Mcp"; "NrdA"; "hypothetical protein SA5_0153/152"; "ORF184" (as bad as gp184); "RNAP1"; "32 kDa protein"; "DUF2732 domain phage protein"; Bad because they don`t mean anything to the casual (or informed) reader.

❑ Do not use the descriptive "putative" **ever**

❑ Unless you are a bioinformatician or biostatistician be very conservative in recording "hits." Could you convince your grandmother (avó)?, if not, list as a "hypothetical protein"

# 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.

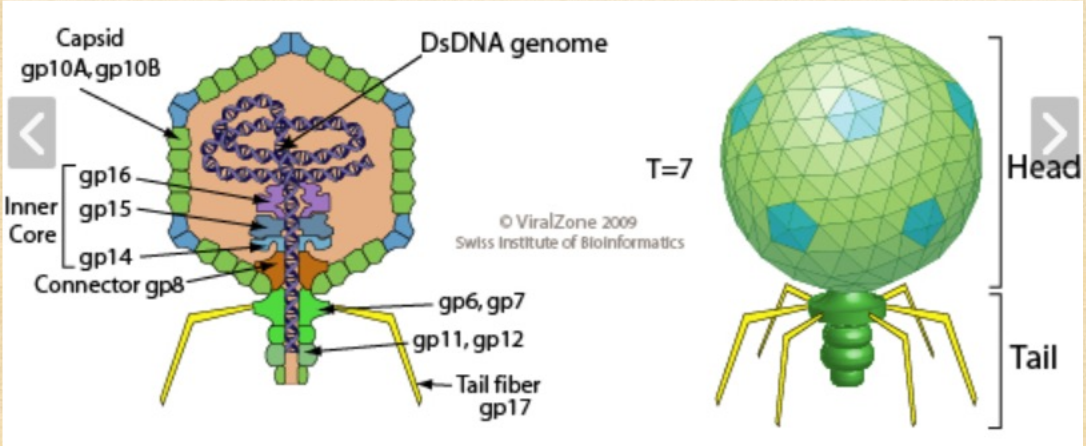


| Entry | Entry name | | Protein names | Gene names | Organism |
|---|---|---|---|---|---|
| P00806 | ENLYS_BPT7 | ☆ | **Endolysin** | 3.5 | Enterobacteria phage T7 T7) |
| P00581 | DPOL_BPT7 | ☆ | **DNA-directed DNA polymerase** | 5 | Enterobacteria phage T7 T7) |
| P03696 | DNBI_BPT7 | ☆ | **Single-stranded DNA-binding protein...** | 2.5 | Enterobacteria phage T7 T7) |
| P03726 | EXLYS_BPT7 | ☆ | **Peptidoglycan transglycosylase gp16** | 16 | Enterobacteria phage T7 T7) |
| P00638 | EXRN_BPT7 | ☆ | **Exonuclease** | 6 | Enterobacteria phage T7 T7) |
| P00969 | DNLI_BPT7 | ☆ | **DNA ligase** | 1.3 | Enterobacteria phage T7 T7) |
| P00641 | ENDO_BPT7 | ☆ | **Endonuclease I** | 3 | Enterobacteria phage T7 T7) |
| P19726 | CAPSA_BPT7 | ☆ | **Major capsid protein** | 10 | Enterobacteria phage T7 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)



# Section 3 – Protein properties

# Protein data extraction from gbk files

- ❑ Sequence Manipulation Suite: GenBank Trans Extractor (http://www.bioinformatics.org/sms2/genbank_trans.html) – may not number the proteins!
- ❑ Genome2D Conversions (http://genome2d.molgenrug.nl/g2d_tools_conversions.html) – choose «Genbank --> Proteins»

# Basic properties of your proteins

- ❑ Number of amino acid residues, mass and pI
- ❑ Sequence Manipulation Suite: Protein Isoelectric Point (http://www.bioinformatics.org/sms2/protein_iep.html)
- ❑ Sequence Manipulation Suite: Protein Molecular Weight (http://www.bioinformatics.org/sms2/protein_mw.html)

# Section 4: Motif searching

# Protein motifs 1

❑You cannot trust BLASTp homolog descriptions

❑Protein motifs:

(a) Batch protein sequence vs profile-HMM database search (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) – offers Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, & TreeFam. Hits should only be considered if E-value ≤ 0.0001

(b) Batch Web-CD Search Tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) adjust E-value to 0.0001

# Protein motifs 2

❑ Protein motifs:

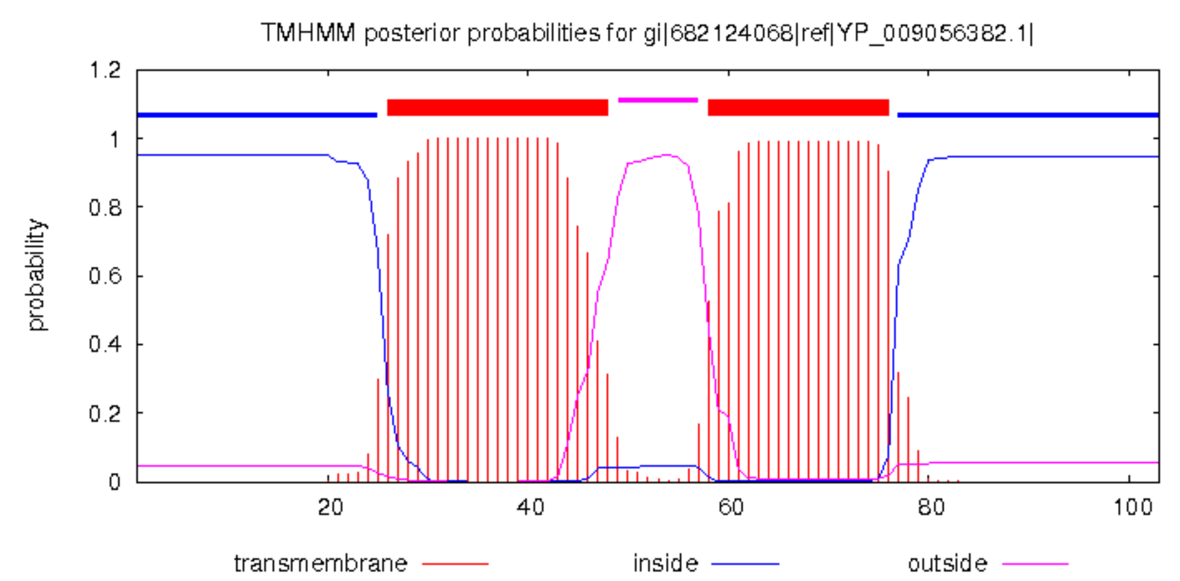
(c) INTERPRO Query Page (http://129.175.105.74/genomics/lbmgeiprscan.html). Unfortunately no E-values for hits

❑Be cautious in interpreting results – employ the grandmother rule

# Protein motifs 3 – TMD 1

- ❑ Transmembrane domains – always use ≥ 2 different servers (chosen from: http://molbiol-tools.ca/Protein_secondary_structure.htm):

  (a) TMHMM

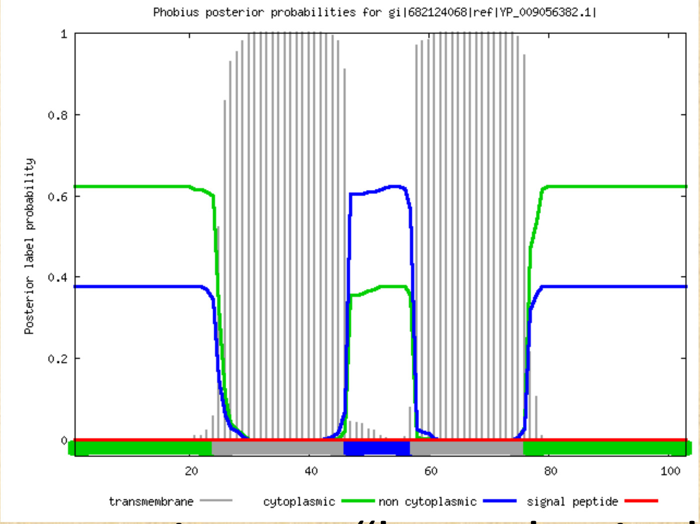


>YP_009056382.1| **holin** [Bacillus phage Bobb]
MENKKETVTQVVEVPTEAPKVEPKMVVLTIVYLVAIINAAAAYLGFDAFNLSVDSERLYEG
VSLFFGVAAFIGAYWKNHDVSKSARIKAAAAKQVDVKQDKVN

# Protein motifs 4 – TMD 2

- ❑ Transmembrane domains – always use ≥ 2 different servers (chosen from: http://molbiol-tools.ca/Protein_secondary_structure.htm):

(b) Phobius



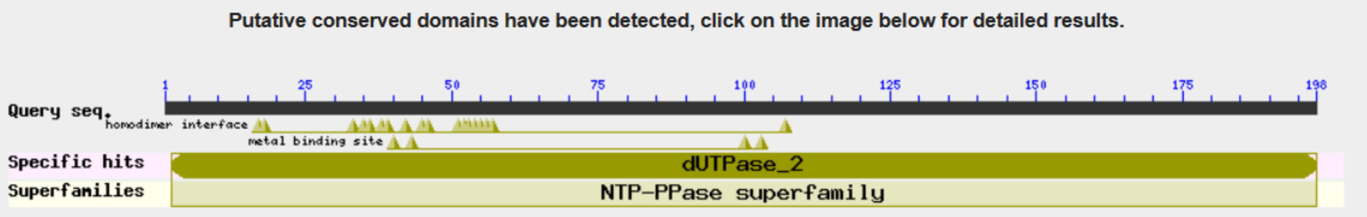
- ❑ If they both agree record the protein as a "hypothetical membrane protein"
- ❑ If the function is know i.e. holin, record data in GenBank file with the following:

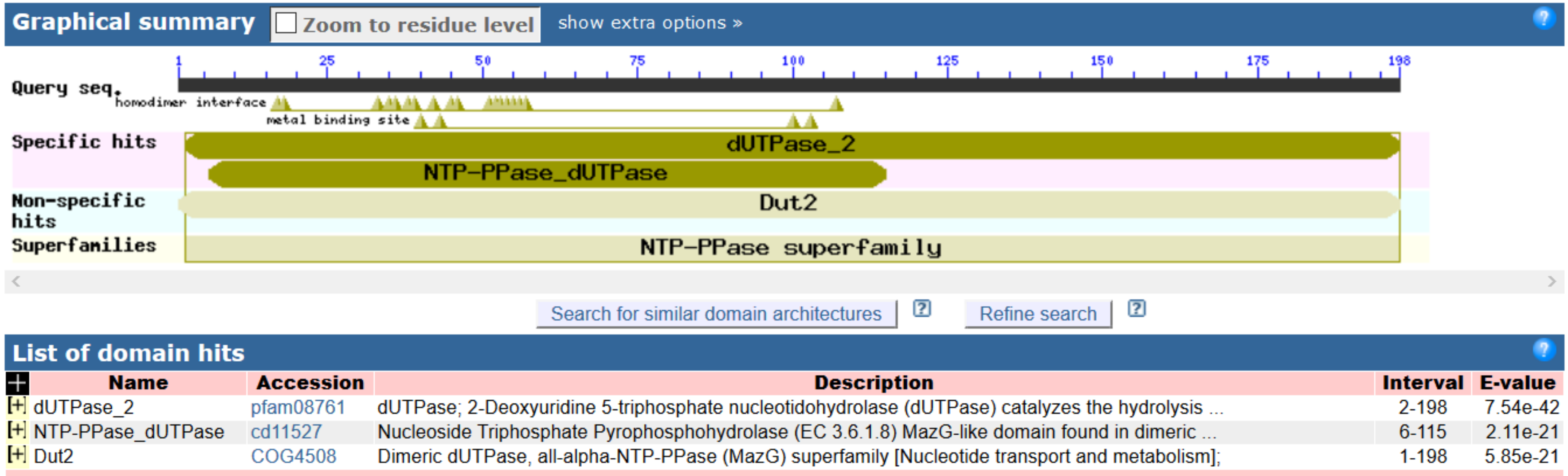
/note="2 transmembrane domains discovered using TMHMM & Phobius"

# Example – *Bacillus* phage dUTPase

>AJK28117.1 dUTPase [Bacillus phage Palmer]
MNLKELFEIQAGLDAEILKNHPIQPGEDRLEKKHAALLVELGEMFNEWRAFKFWSHDKEPRMAVKCPECEGAAARQASDGSYVECGTCDGAGTIDKVL
KELVDCLHFVLSIGLEHEFDTKLNMVIEPILFSRSDDGNNIIAQFIELLKVEWELVGRHYKEGLELFIGFCEMLGYTWEQVREAYLIKNQENHYRQMNGY

- ❑BLASTp vs nr and Viruses (taxid:10239) databases – motif “hits”


Putative conserved domains have been detected, click on the image below for detailed results.



List of domain hits

| | Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|---|
| [+] | dUTPase_2 | pfam08761 | dUTPase; 2-Deoxyuridine 5-triphosphate nucleotidohydrolase (dUTPase) catalyzes the hydrolysis ... | 2-198 | 7.54e-42 |
| [+] | NTP-PPase_dUTPase | cd11527 | Nucleoside Triphosphate Pyrophosphohydrolase (EC 3.6.1.8) MazG-like domain found in dimeric ... | 6-115 | 2.11e-21 |
| [+] | Dut2 | COG4508 | Dimeric dUTPase, all-alpha-NTP-PPase (MazG) superfamily [Nucleotide transport and metabolism]; | 1-198 | 5.85e-21 |

- ❑Low E-value hits to three motif databases

# HHpred - Homology detection & structure prediction by HMM-HMM comparison

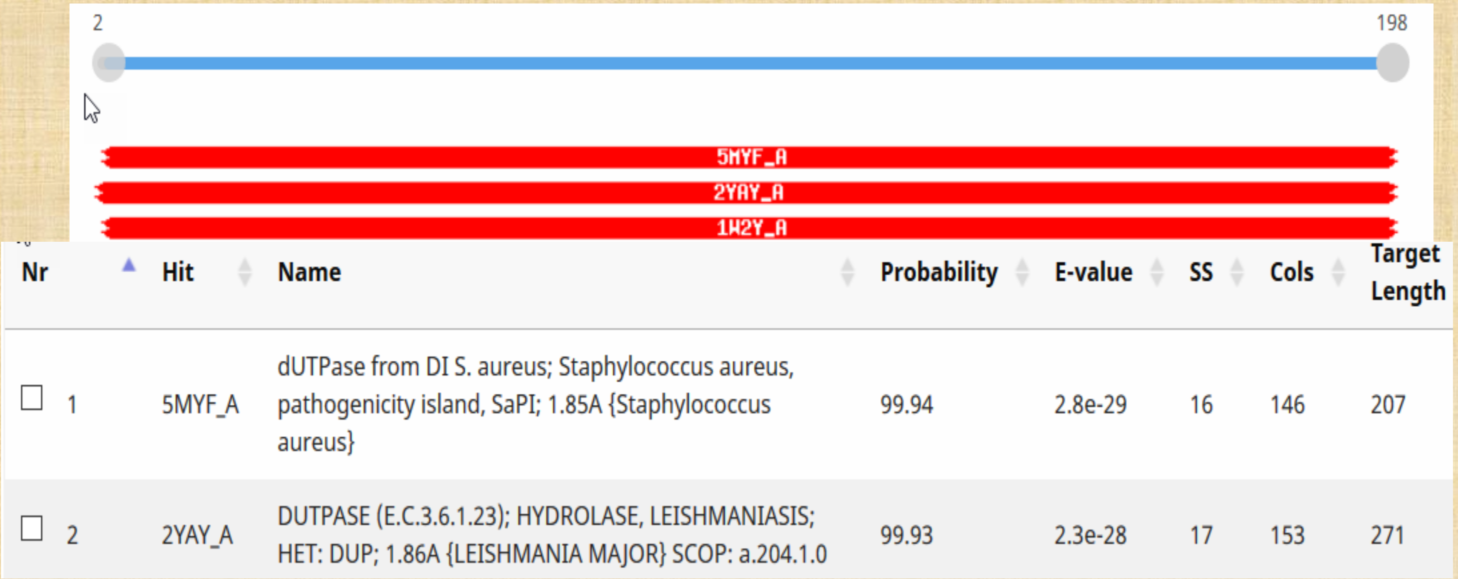
"It is well known that sequence search methods such as BLAST, FASTA, or PSI-BLAST are of prime importance for biological research because functional information of a protein or gene can be inferred from homologous proteins or genes identified in a sequence search. But quite often no significant relationship to a protein of known function can be established.

It is less well known that in cases where conventional sequence search methods fail, the recently developed, highly sensitive methods for homology detection or structure prediction quite often allow to one to make inferences from more remotely homologous relationships."

- ❑ https://toolkit.tuebingen.mpg.de/#/tools/hhpred
- ❑ Single protein, no batch mode unless you download program & database
- ❑ Retain information if "Prob" is ≥ 90% & hit is to phage protein

# Example – *Bacillus* phage dUTPase 2

- ❑HHpred analysis



| Nr | Hit | Name | Probability | E-value | SS | Cols | Target Length |
|---|---|---|---|---|---|---|---|
| □ 1 | 5MYF_A | dUTPase from DI S. aureus; Staphylococcus aureus, pathogenicity island, SaPI; 1.85A {Staphylococcus aureus} | 99.94 | 2.8e-29 | 16 | 146 | 207 |
| □ 2 | 2YAY_A | DUTPASE (E.C.3.6.1.23); HYDROLASE, LEISHMANIASIS; HET: DUP; 1.86A {LEISHMANIA MAJOR} SCOP: a.204.1.0 | 99.93 | 2.3e-28 | 17 | 153 | 271 |

- ❑High scoring “hits” to proteins all called dUTPases
- ❑5MYF can be visualized at NCBI (https://www.ncbi.nlm.nih.gov/Structure/) or RCSB PDB (https://www.rcsb.org/)

# Bottom line

- ❑ Good evidence here that this protein is a deoxyuridine triphosphatase (dUTPase)
- ❑ But, if you couldn’t convince your grandmother that a protein is a “dUTPase” describe it as a “hypothetical protein”

# Questions?

# Intro to classification & taxonomy

Evelien Adriaenssens

evelien.adriaenssens@quadram.ac.uk

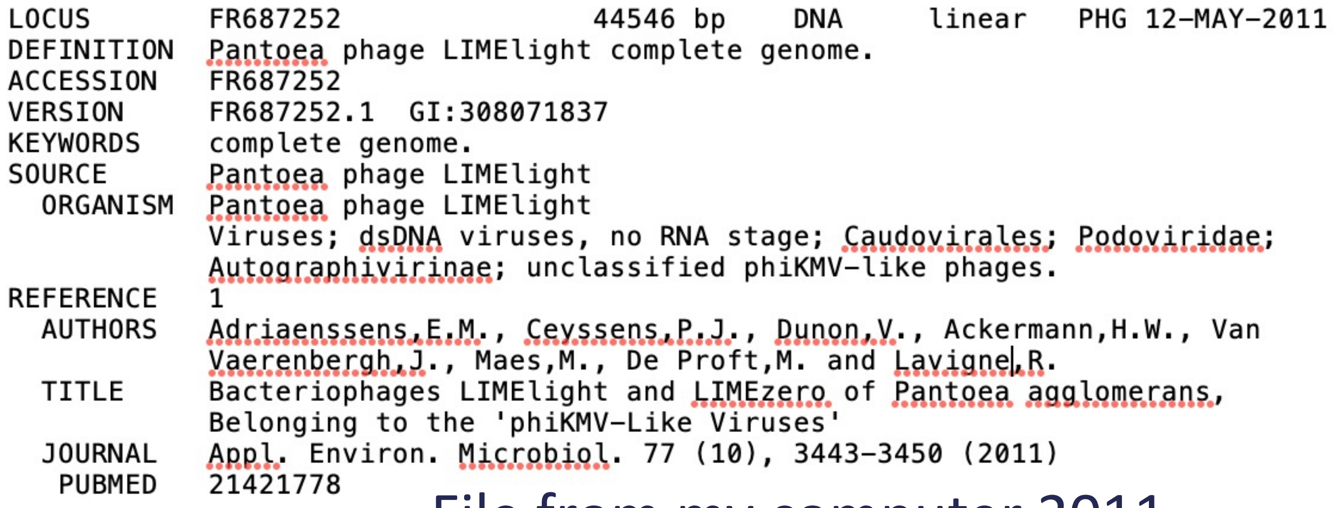
# Aim

- Provide you with the information and tools to fill in the <ORGANISM> section of a GenBank file

→ Gets automatically updated after taxonomy updates

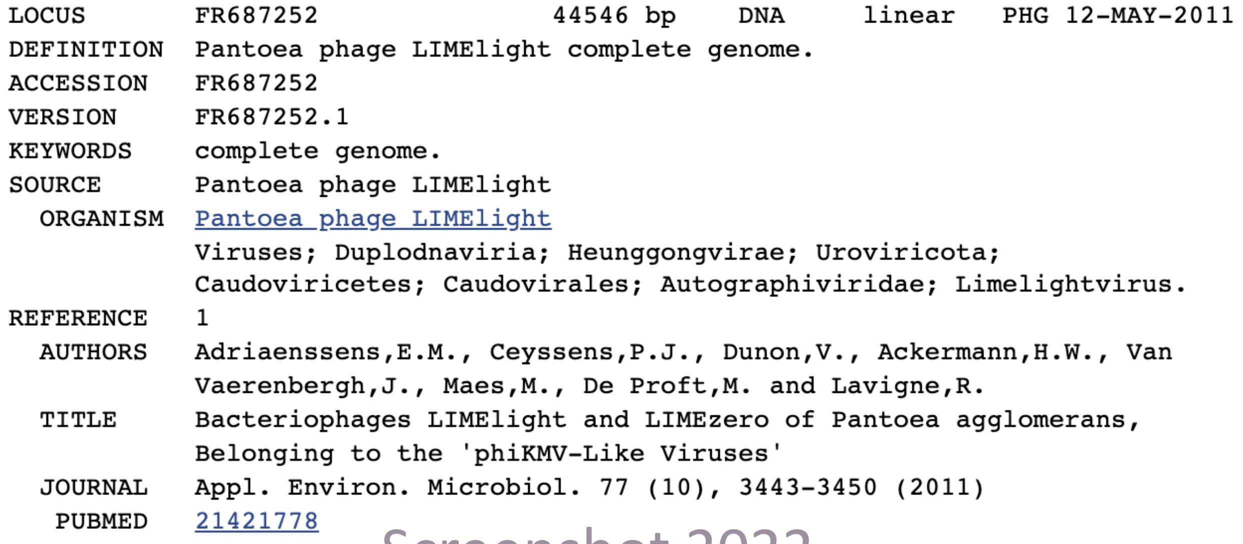
→ Fill in lineage to closest available taxon and then add ”unclassified”

→ Don’t use taxonomy information in the phage name! (DEFINITION field)



```
LOCUS       FR687252              44546 bp    DNA     linear   PHG 12-MAY-2011
DEFINITION  Pantoea phage LIMElight complete genome.
ACCESSION   FR687252
VERSION     FR687252.1  GI:308071837
KEYWORDS    complete genome.
SOURCE      Pantoea phage LIMElight
  ORGANISM  Pantoea phage LIMElight
            Viruses; dsDNA viruses, no RNA stage; Caudovirales; Podoviridae;
            Autographivirinae; unclassified phiKMV-like phages.
REFERENCE   1
  AUTHORS   Adriaenssens,E.M., Ceyssens,P.J., Dunon,V., Ackermann,H.W., Van
            Vaerenbergh,J., Maes,M., De Proft,M. and Lavigne,R.
  TITLE     Bacteriophages LIMElight and LIMEzero of Pantoea agglomerans,
            Belonging to the 'phiKMV-Like Viruses'
  JOURNAL   Appl. Environ. Microbiol. 77 (10), 3443-3450 (2011)
   PUBMED   21421778
```

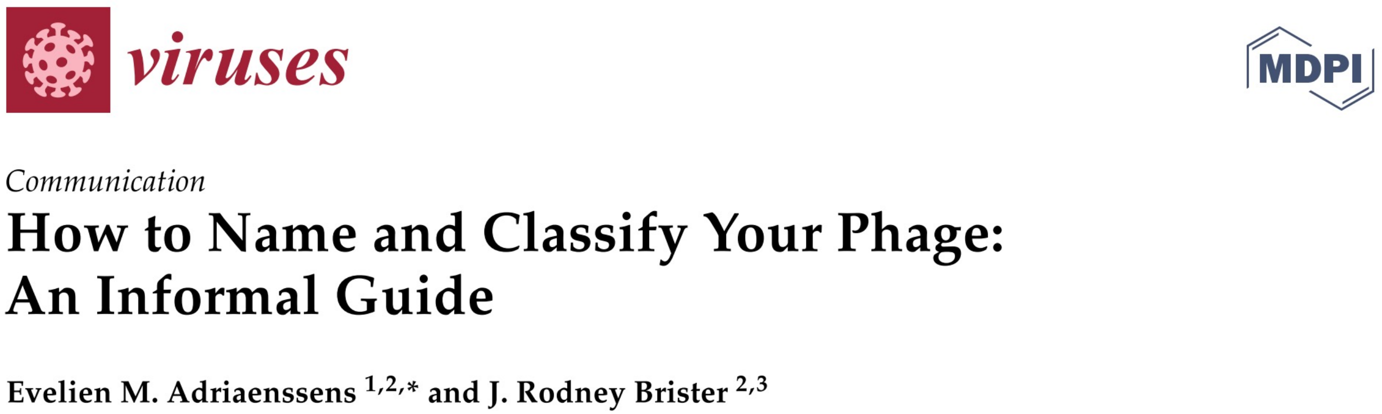
File from my computer 2011



```
LOCUS       FR687252              44546 bp    DNA     linear   PHG 12-MAY-2011
DEFINITION  Pantoea phage LIMElight complete genome.
ACCESSION   FR687252
VERSION     FR687252.1
KEYWORDS    complete genome.
SOURCE      Pantoea phage LIMElight
  ORGANISM  Pantoea phage LIMElight
            Viruses; Duplodnaviria; Heunggongvirae; Uroviricota;
            Caudoviricetes; Caudovirales; Autographiviridae; Limelightvirus.
REFERENCE   1
  AUTHORS   Adriaenssens,E.M., Ceyssens,P.J., Dunon,V., Ackermann,H.W., Van
            Vaerenbergh,J., Maes,M., De Proft,M. and Lavigne,R.
  TITLE     Bacteriophages LIMElight and LIMEzero of Pantoea agglomerans,
            Belonging to the 'phiKMV-Like Viruses'
  JOURNAL   Appl. Environ. Microbiol. 77 (10), 3443-3450 (2011)
   PUBMED   21421778
```

Screenshot 2022

# Recent resources



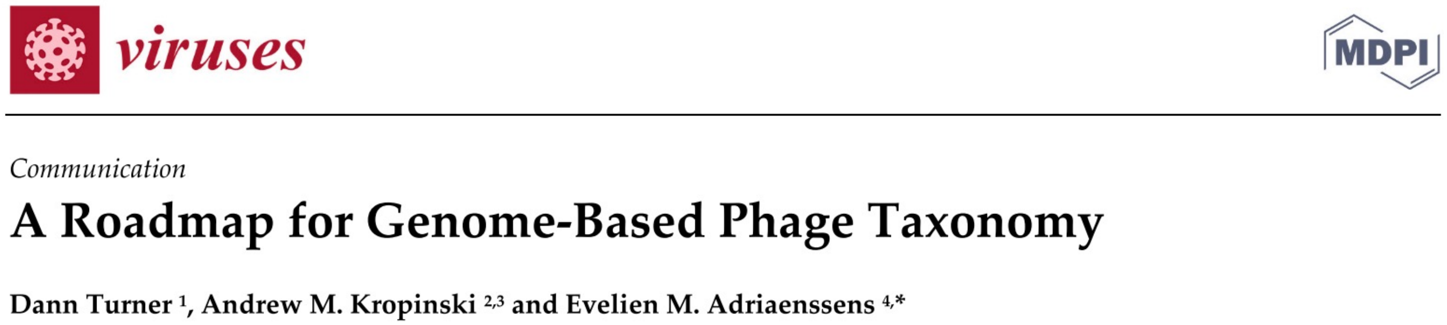
viruses

MDPI

*Communication*

## How to Name and Classify Your Phage: An Informal Guide

Evelien M. Adriaenssens [1,2,*] and J. Rodney Brister [2,3]



viruses

MDPI

*Communication*

## A Roadmap for Genome-Based Phage Taxonomy

Dann Turner [1], Andrew M. Kropinski [2,3] and Evelien M. Adriaenssens [4,*]

# Naming your phage

- No official rules about naming phage/virus isolates
- BUT lots of rules for official taxon names (e.g. no hyphens or slashes, no Greek letters...)
- **BE UNIQUE!**
- ICTV BVS has used the exemplar isolate name as basis for the species and/or genus names in the past

**Remember: species != phage**

all domestic dogs member of the species *Canis lupus*

# Binomial species naming system

Use genus name plus species epithet to refer to virus species in freeform format

**Examples:**
Salmonella phage P22, member of genus *Lederbergvirus,* exemplar isolate of species *Lederbergvirus P22*
Enterobacteria phage MS2, member of genus *Emesvirus*, exemplar isolate of species *Emesvirus zinderi*

Clear difference between phage isolate and species!

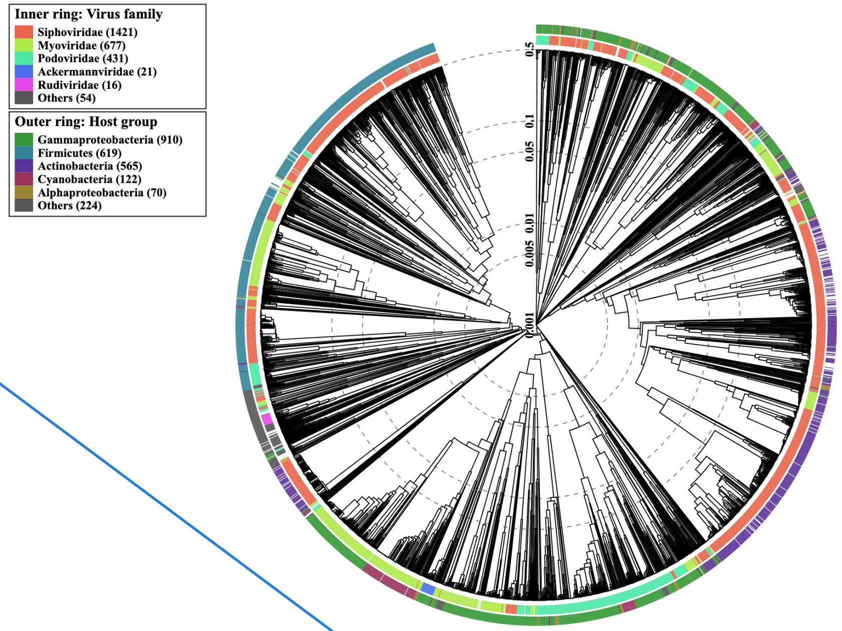
**In practice:** my phage is called Salmonella phage Tweedledum and it belongs to the species *Lederbergvirus P22*.

# Basic phage classification workflow

**Start:** well-annotated phage

Find database relatives

BLAST, HMMs, VIPtree, GRAViTy vConTACT2



Use all information collected along the way!

Determine the intergenomic distance

VIRIDIC, pyANI, CD-HIT-EST…

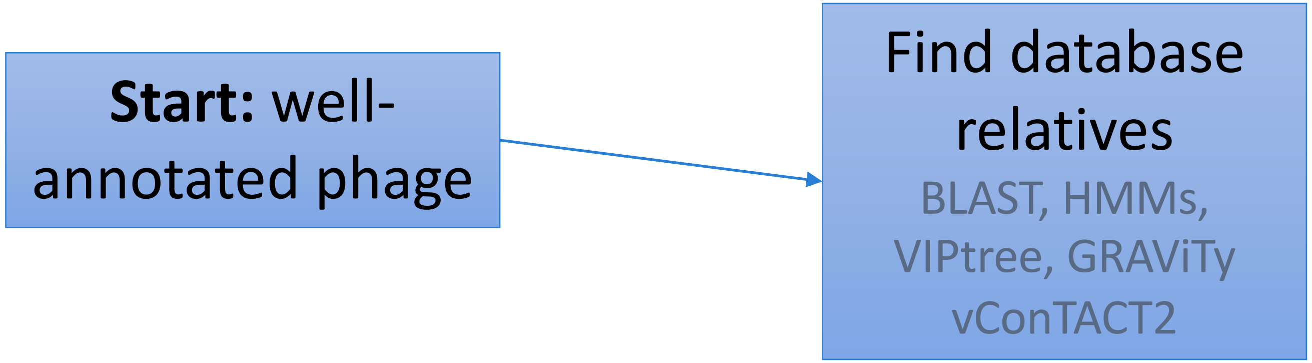
Determine shared protein content

CoreGenes 5.0; GET_HOMOLOGUES, OrthoMCL…

Multiple sequence alignment & phylogenetics of signature genes

ClustalΩ, MAFFT, MUSCLE, Phylogeny.fr, IQ-Tree, raxML, FastTree…

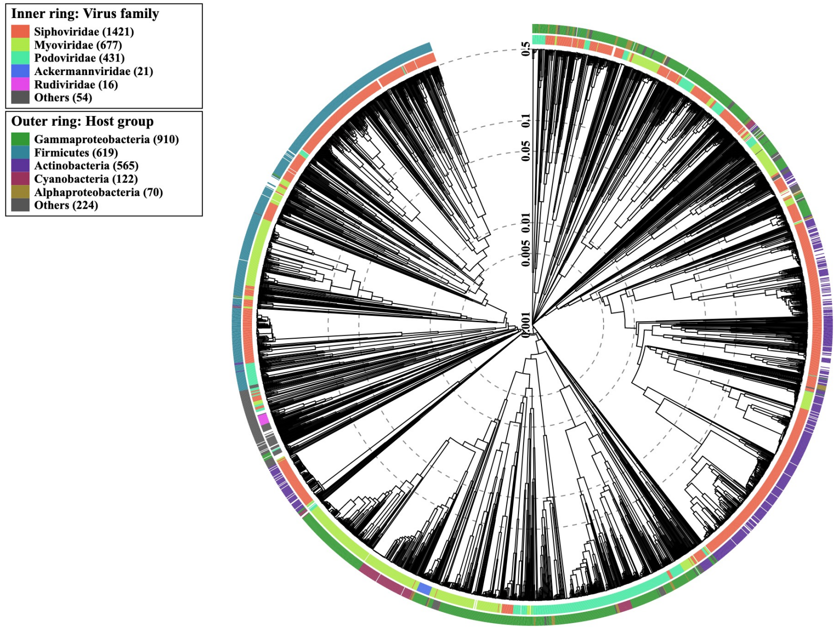
# Basic phage classification workflow



**Start:** well-annotated phage

Find database relatives

BLAST, HMMs, VIPtree, GRAViTy vConTACT2



Step 1: find relatives
How closely related are they?

# Using BLAST

- **BLASTn:** compare genome to genome

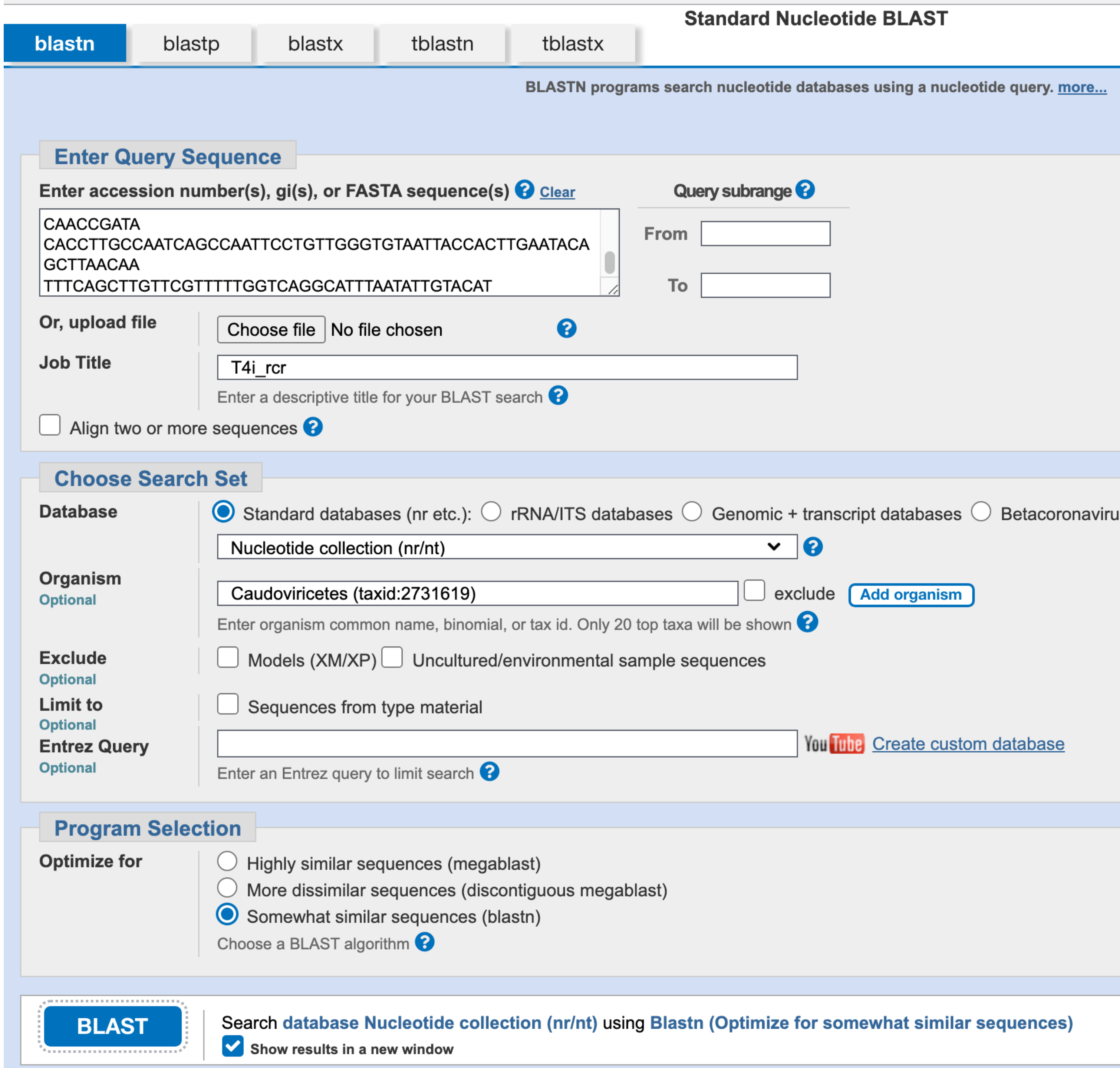
→ Limit search to subset of organisms (eg. viruses or *Caudoviricetes*)

→ Use “somewhat similar sequences” first

- **BLASTx:** compare genome to protein database
- → If BLASTn doesn’t yield a result

- **tBLASTx:** compare translated genome with translate genome

→ Very computationally demanding, not recommended online



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

## Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

```
CAACCGATA
CACCTTGCCAATCAGCCAATTCCTGTTGGGTGTAATTACCACTTGAATACA
GCTTAACAA
TTTCAGCTTGTTCGTTTTTGGTCAGGCATTTAATATTGTACAT
```

Query subrange ?

From

To

Or, upload file: Choose file No file chosen ?

Job Title: T4i_rcr

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

## Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronaviru

Nucleotide collection (nr/nt) ?

Organism Optional: Caudoviricetes (taxid:2731619) □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search ?

## Program Selection

Optimize for:
○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

BLAST

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

☑ Show results in a new window

# Alternative online location to start BLAST: NCBI Virus

- https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

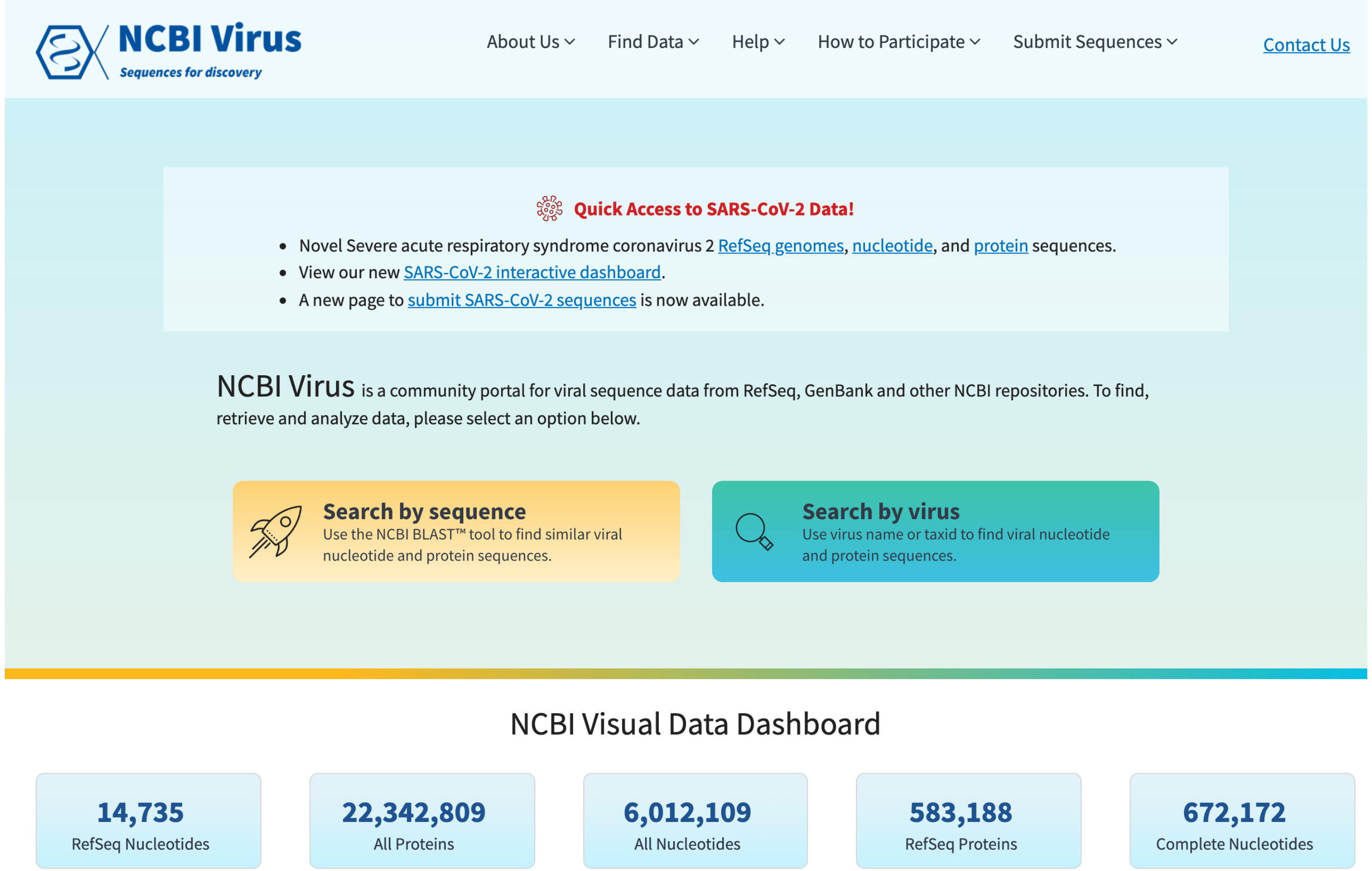
- Automatically limited to virus database

→ Easy refinement of search results

→ Extensive metadata in tabular form

→ More detailed investigation possible of search results

→ Easy download of selected search results



Quick Access to SARS-CoV-2 Data!

- Novel Severe acute respiratory syndrome coronavirus 2 RefSeq genomes, nucleotide, and protein sequences.
- View our new SARS-CoV-2 interactive dashboard.
- A new page to submit SARS-CoV-2 sequences is now available.

NCBI Virus is a community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below.

Search by sequence
Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.
Search by virus
Use virus name or taxid to find viral nucleotide and protein sequences.

NCBI Visual Data Dashboard

| 14,735 | 22,342,809 | 6,012,109 | 583,188 | 672,172 |
|---|---|---|---|---|
| RefSeq Nucleotides | All Proteins | All Nucleotides | RefSeq Proteins | Complete Nucleotides |

# Using VipTree to situate new phage genome

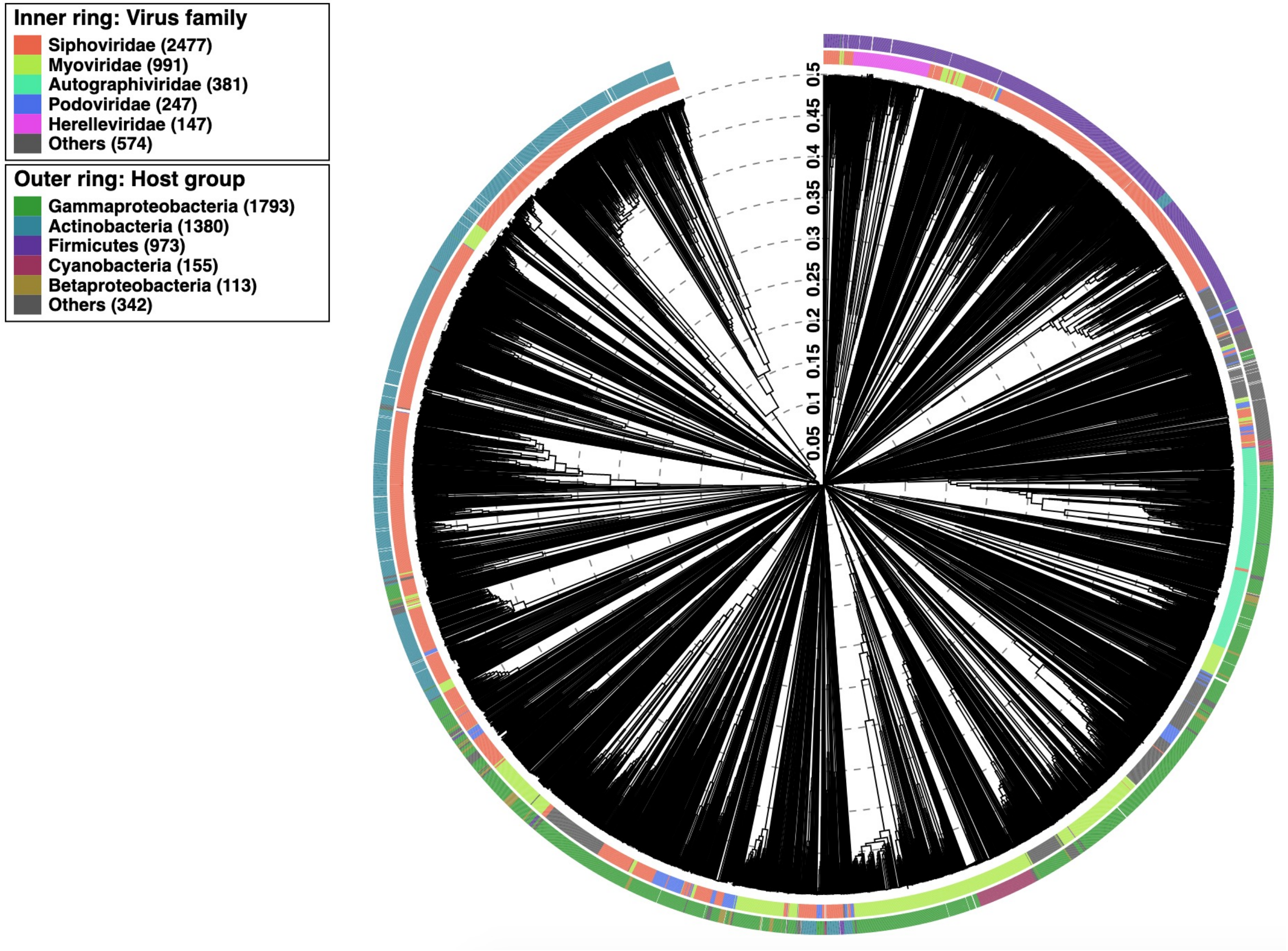
Based on phage proteomic tree approach

Different trees for different virus types

Can upload up to 100 genomes

Branch lengths scaled from 0 to 0.5 (0 identical at amino acid level, 0.5 no similarity)

Taxonomy not up to date

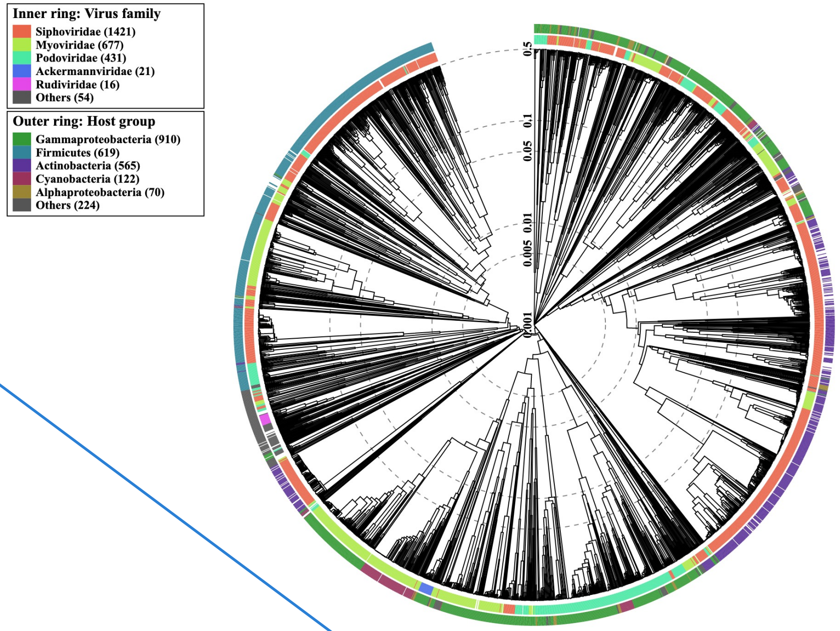


# Basic phage classification workflow

**Start:** well-annotated phage

Find database relatives

BLAST, HMMs, VIPtree, GRAViTy vConTACT2



Determine the intergenomic distance

VIRIDIC, pyANI, CD-HIT-EST…

# Does my new phage represent a new species?

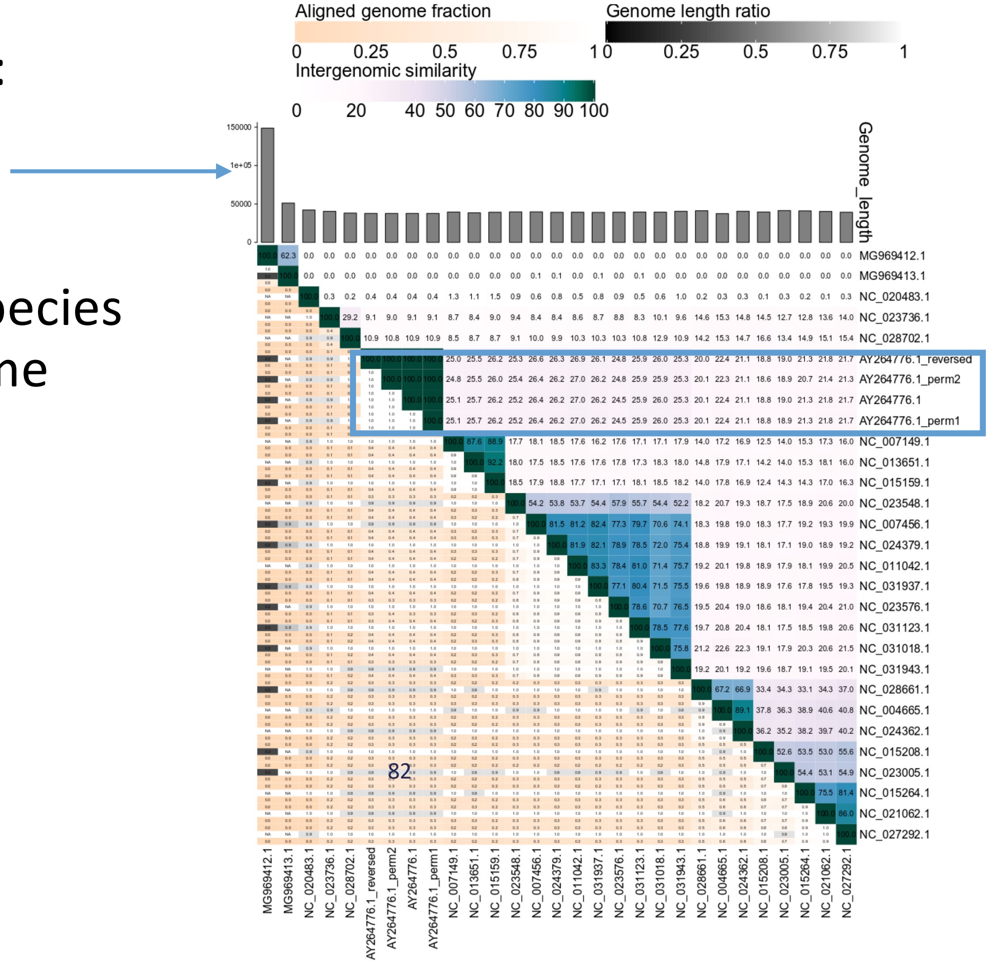
- Main species demarcation criterion for bacteriophages: genome sequence identity of 95%

→ the genomes of two isolates belonging to the same species differ from each other by less than 5% over the genome length

→ Suggested tool to use: VIRIDIC (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/)

→ check for synteny, isolates with high levels of rearrangements do not belong to same species

➔ part of existing species: use this taxonomic description to deposit in GenBank/EMBL/DDBJ



VIRIDIC example, Moraru et al 2020, Viruses

# Does my phage belong to a new genus?

Genus: cohesive group of viruses sharing a high degree of nucleotide sequence similarity (generally > 70%), monophyletic group in marker gene phylogenetic tree

Other potential defining characteristics:

- average genome length
- average number of CDS
- percentage of shared CDS
- genome organisation
- presence of tRNAs
- presence of certain signature genes

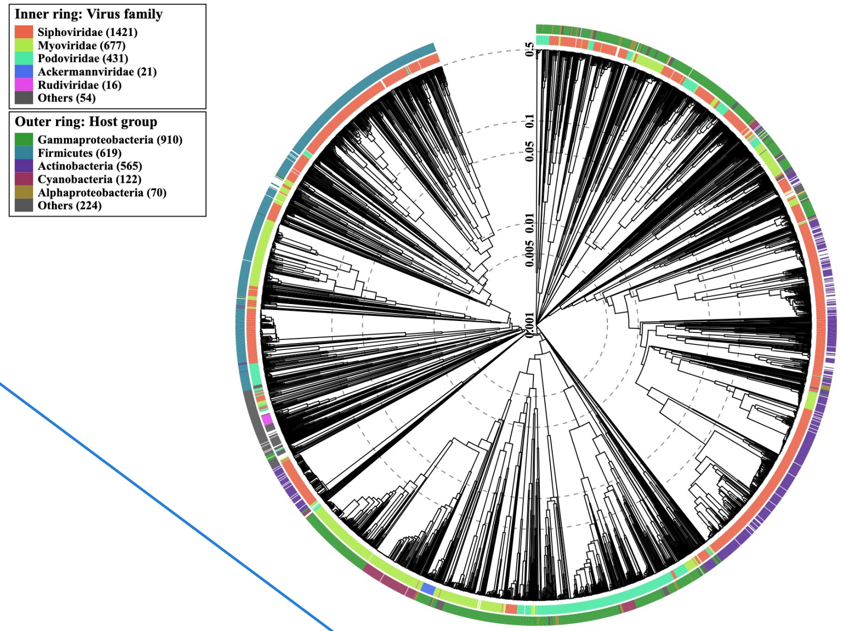
➔ New genus: submit taxonomy proposal with Chair of Subcommittee, or Study Group Chair

# Basic phage classification workflow

**Start:** well-annotated phage

Find database relatives

BLAST, HMMs, VIPtree, GRAViTy vConTACT2



Use all information collected along the way!

Determine the intergenomic distance

VIRIDIC, pyANI, CD-HIT-EST…

Determine shared protein content

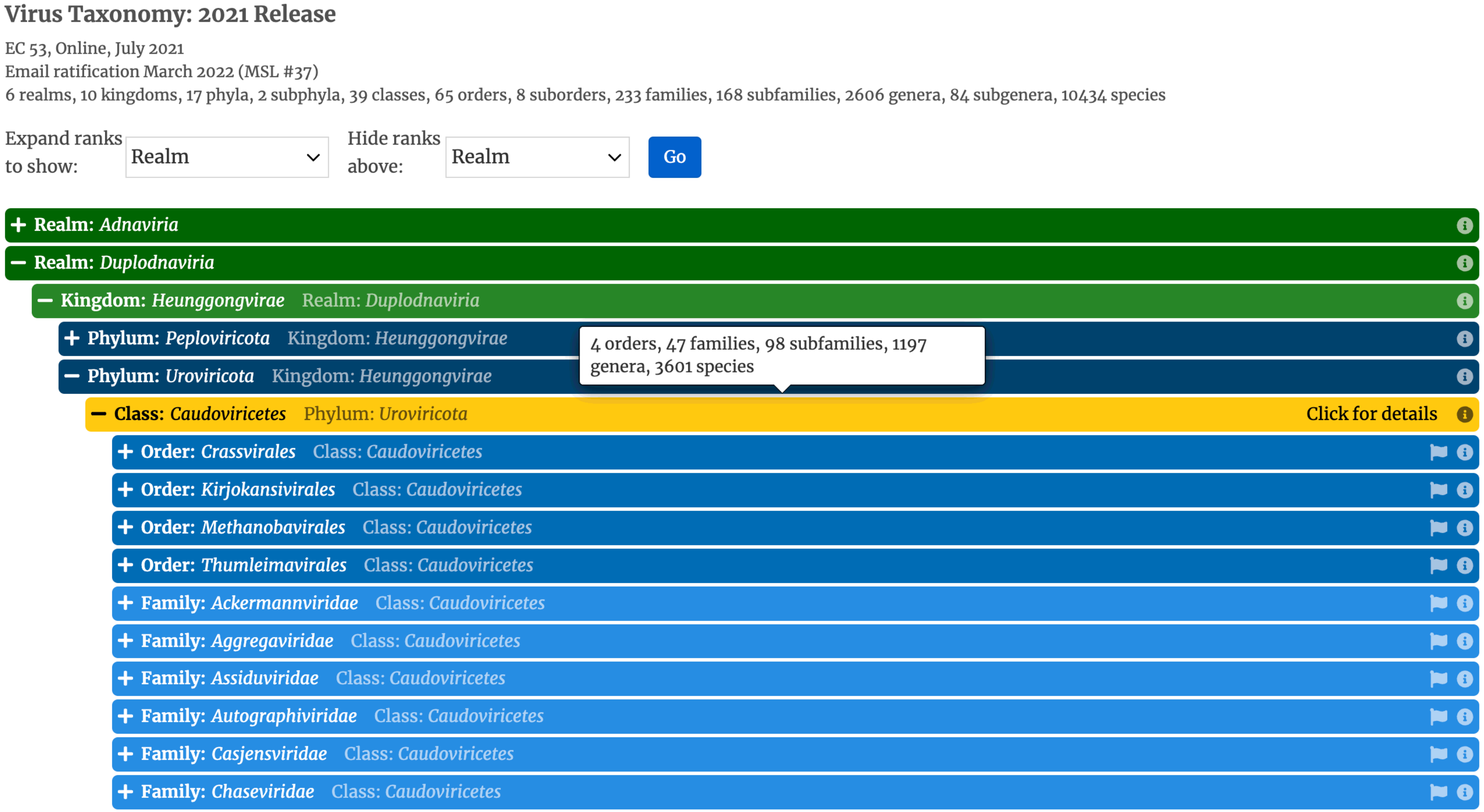
CoreGenes 5.0; GET_HOMOLOGUES, OrthoMCL…

Multiple sequence alignment & phylogenetics of signature genes

ClustalΩ, MAFFT, MUSCLE, Phylogeny.fr, IQ-Tree, raxML, FastTree…

# Does my phage belong to an existing subfamily & family?

- Assessed with a combination of genomic, proteomic and phylogenetic tools
- Check demarcation criteria for families: https://ictv.global/taxonomy



**Virus Taxonomy: 2021 Release**

EC 53, Online, July 2021
Email ratification March 2022 (MSL #37)
6 realms, 10 kingdoms, 17 phyla, 2 subphyla, 39 classes, 65 orders, 8 suborders, 233 families, 168 subfamilies, 2606 genera, 84 subgenera, 10434 species

Expand ranks to show: Realm

Hide ranks above: Realm

Go

+ **Realm:** *Adnaviria*
− **Realm:** *Duplodnaviria*
− **Kingdom:** *Heunggongvirae* Realm: *Duplodnaviria*
+ **Phylum:** *Peploviricota* Kingdom: *Heunggongvirae*
− **Phylum:** *Uroviricota* Kingdom: *Heunggongvirae*

4 orders, 47 families, 98 subfamilies, 1197 genera, 3601 species

− **Class:** *Caudoviricetes* Phylum: *Uroviricota* Click for details
+ **Order:** *Crassvirales* Class: *Caudoviricetes*
+ **Order:** *Kirjokansivirales* Class: *Caudoviricetes*
+ **Order:** *Methanobavirales* Class: *Caudoviricetes*
+ **Order:** *Thumleimavirales* Class: *Caudoviricetes*
+ **Family:** *Ackermannviridae* Class: *Caudoviricetes*
+ **Family:** *Aggregaviridae* Class: *Caudoviricetes*
+ **Family:** *Assiduviridae* Class: *Caudoviricetes*
+ **Family:** *Autographiviridae* Class: *Caudoviricetes*
+ **Family:** *Casjensviridae* Class: *Caudoviricetes*
+ **Family:** *Chaseviridae* Class: *Caudoviricetes*

Hover over for more information

Click for details will show the taxonomy proposals:
- demarcation criteria
- marker genes

# New subfamily & family?

- Advanced taxonomy
- Contact members of the Bacterial Viruses Subcommittee: https://ictv.global/sc/bacterial

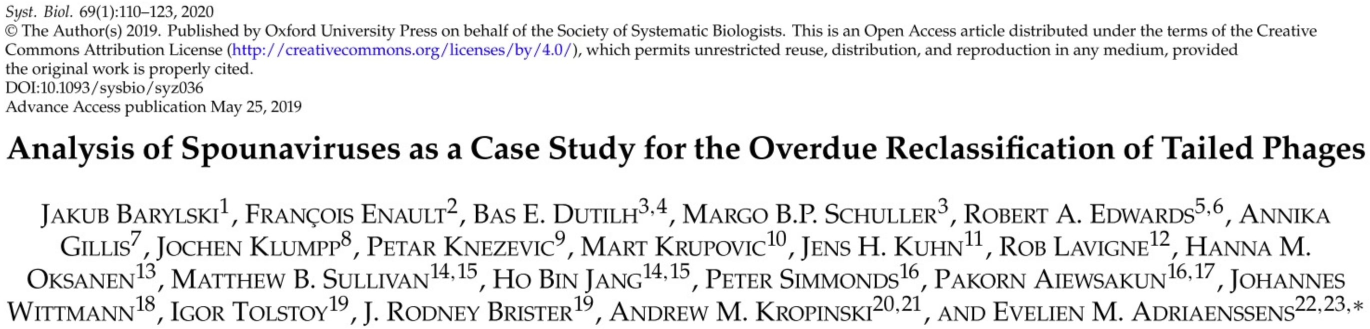
Examples of creating new families:

*Herelleviridae*

https://academic.oup.com/sysbio/article/69/1/110/5498714

*Schitoviridae*

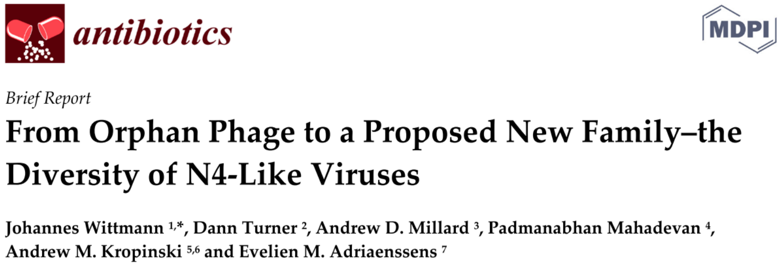
https://www.mdpi.com/2079-6382/9/10/663



*Syst. Biol.* 69(1):110–123, 2020
© The Author(s) 2019. Published by Oxford University Press on behalf of the Society of Systematic Biologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syz036
Advance Access publication May 25, 2019

**Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages**

Jakub Barylski[1], François Enault[2], Bas E. Dutilh[3,4], Margo B.P. Schuller[3], Robert A. Edwards[5,6], Annika Gillis[7], Jochen Klumpp[8], Petar Knezevic[9], Mart Krupovic[10], Jens H. Kuhn[11], Rob Lavigne[12], Hanna M. Oksanen[13], Matthew B. Sullivan[14,15], Ho Bin Jang[14,15], Peter Simmonds[16], Pakorn Aiewsakun[16,17], Johannes Wittmann[18], Igor Tolstoy[19], J. Rodney Brister[19], Andrew M. Kropinski[20,21], and Evelien M. Adriaenssens[22,23,*]



antibiotics

MDPI

*Brief Report*

**From Orphan Phage to a Proposed New Family–the Diversity of N4-Like Viruses**

Johannes Wittmann [1,*], Dann Turner [2], Andrew D. Millard [3], Padmanabhan Mahadevan [4], Andrew M. Kropinski [5,6] and Evelien M. Adriaenssens [7]

# Identify the Core Genome for a family

- Number of shared genes will depend on genome size of new family

- Webserver: CoreGenes 5.0 https://coregenes.ngrok.io/

- Command line tools for (bacterial) pangenomics analyses can also be used.
- GET_HOMOLOGUES
- Roary
- PIRATE
- OrthoMCL

→ Advanced classification, not the scope of this workshop

# In summary

To classify a phage:

- Find relatives in public databases
- Identify the relationships at the nucleotide level
- Identify the relationships at the predicted proteome level
- Perform phylogenetics (or phylogenomics)
- **Submit a Taxonomy Proposal to Study Group Chair or Subcommittee Chair (Evelien)**

# Submission to INSDC

- Different workflows for GenBank, ENA and DDBJ
  - GenBank: https://www.ncbi.nlm.nih.gov/books/NBK566995/
    - BankIt: https://www.ncbi.nlm.nih.gov/WebSub/html/requirements.html
    - https://www.ncbi.nlm.nih.gov/WebSub/html/help/feature-table.html

- ENA: https://ena-docs.readthedocs.io/en/latest/submit/general-guide/interactive.html

- https://www.ddbj.nig.ac.jp/ddbj/submission-e.html

ICTV

UWE Bristol

UNIVERSITY of GUELPH

Quadram Institute
Science Health Food Innovation

UKRI Biotechnology and Biological Sciences Research Council

UKRI Medical Research Council