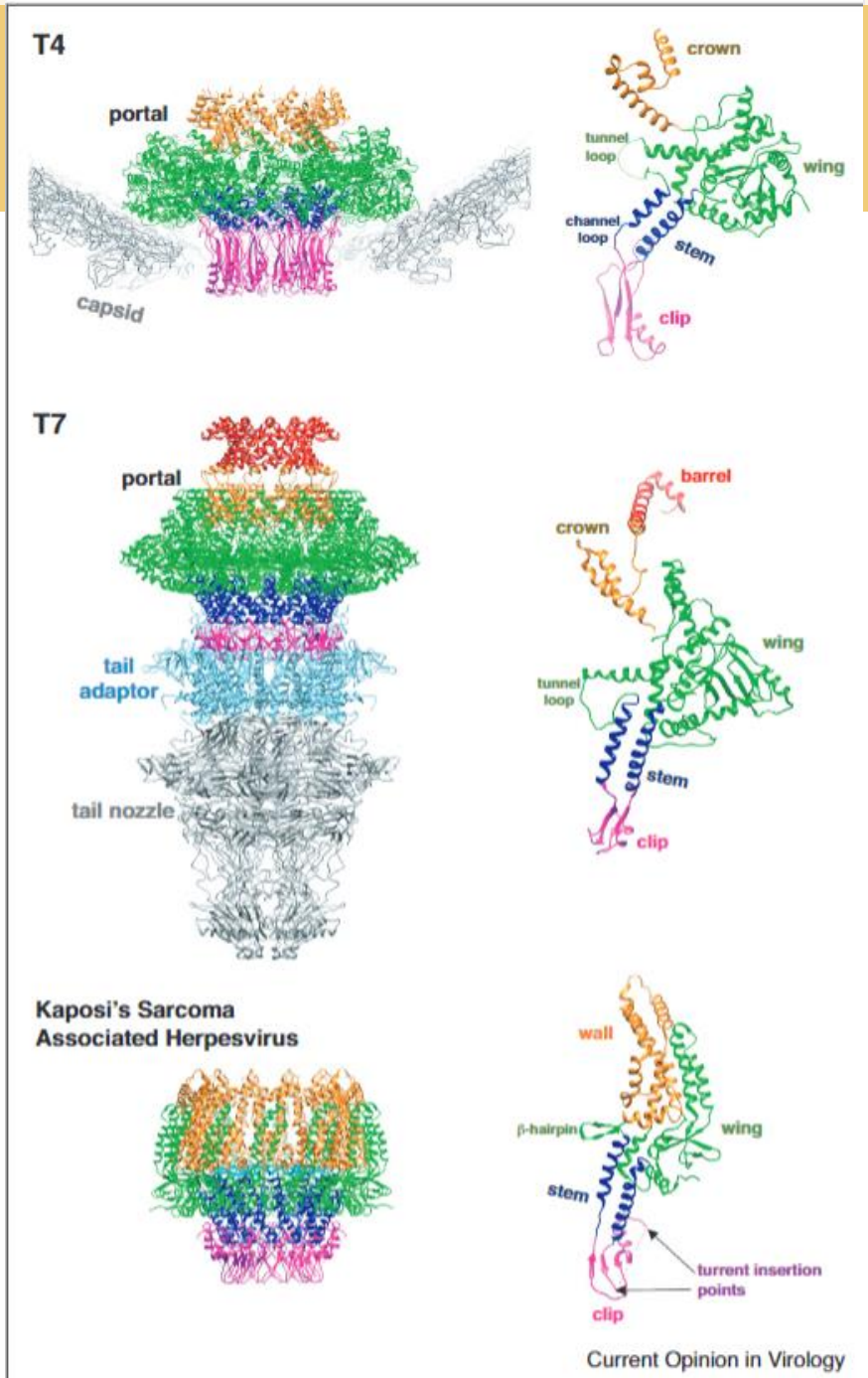
MGV Database: Extracting Connector Protein

Goal

1. Find connector protein sequence
2. Simulate sequences

Improve understanding of the general connector protein structure

Eventual goal: We want to simulate the structure of an entire virus



Viral Protein Homology

- Connector protein sequences aren’t similar because the phages evolved the protein independently
- Common methods of annotating proteins:
    - Structural: modelling and comparing
    - Sequence: comparing sequence to database
- Homology = sequence related because of common ancestor
- Similarity = degree of likeness between sequences

BUT the genomic organization (order of genes) and structure is similar

Phages have adapted the same protein fold to fulfill multiple functions in virion assembly

Lia Cardarelli[a], Lisa G. Pell[a,b], Philipp Neudecker[a,c,d], Nawaz Pirani[a,b], Amanda Liu[c], Lindsay A. Baker[a,b], John L. Rubinstein[a,b], Karen L. Maxwell[c], and Alan R. Davidson[a,c,1]

Departments of [a]Biochemistry and [c]Molecular Genetics, University of Toronto, Toronto, ON, Canada M5S 1A8; [b]Molecular Structure and Function Program, The Hospital for Sick Children Research Institute, Toronto, ON, Canada M5G 1X8; and [d]Department of Chemistry, University of Toronto, Toronto, ON, Canada M5S 3H6

Edited* by Michael G. Rossmann, Purdue University, West Lafayette, IN, and approved June 21, 2010 (received for review April 28, 2010)

Evolutionary relationships may exist among very diverse groups of proteins even though they perform different functions and display little sequence similarity. The tailed bacteriophages present a uniquely amenable system for identifying such groups because of their huge diversity yet conserved genome structures. In this work, we used structural, functional, and genomic context comparisons to conclude that the head–tail connector protein and tail tube protein of bacteriophage λ diverged from a common ancestral protein. Further comparisons of tertiary and quaternary structures indicate that the baseplate hub and tail terminator proteins of bacteriophage may also be part of this same family. We propose that all of these proteins evolved from a single ancestral tail tube protein fold, and that gene duplication followed by differentiation

the connector and passes down the tail into the cell. The portion of the connector that is inserted into the head is composed of a dodecameric ring of the product of gene *B* (gpB), also known as the portal protein. The bottom surface of the connector (Fig. 1*A*), which interacts with the tail, is composed of gpFII (5). Another protein, gpW, is required for the stabilization of the DNA within the head and for the addition of gpFII (6, 7), suggesting that it may be positioned in the connector between gpB and gpFII. *Bacillus subtilis* phage SPP1 gp16, a protein with the same structure, function, and genomic position as gpFII (2) (Fig. 1 *A* and *C*), has been shown by cryoelectron microscopy (cryoEM) to form a 12-membered ring within the connector (8, 9). Although the number of molecules of gpFII in assembled phage particles has

Data

Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome

Stephen Nayfach[1,2], David Páez-Espino[1,2], Lee Call[1,2], Soo Jen Low[3], Hila Sberro[4,5], Natalia N. Ivanova[1,2], Amy D. Proal[6], Michael A. Fischbach[7,8,9,10], Ami S. Bhatt[4,5], Philip Hugenholtz[3] and Nikos C. Kyrpides[1,2]

Bacteriophages have important roles in the ecology of the human gut microbiome but are under-represented in reference databases. To address this problem, we assembled the Metagenomic Gut Virus catalogue that comprises 189,680 viral genomes from 11,810 publicly available human stool metagenomes. Over 75% of genomes represent double-stranded DNA phages that infect members of the Bacteroidia and Clostridia classes. Based on sequence clustering we identified 54,118 candidate viral species, 92% of which were not found in existing databases. The Metagenomic Gut Virus catalogue improves detection of viruses in stool metagenomes and accounts for nearly 40% of CRISPR spacers found in human gut Bacteria and Archaea. We also produced a catalogue of 459,375 viral protein clusters to explore the functional potential of the gut virome. This revealed tens of thousands of diversity-generating retroelements, which use error-prone reverse transcription to mutate target genes and may be involved in the molecular arms race between phages and their bacterial hosts.

- Paper analyzed bacteriophages found in the human gut microbiome
- 189,680 viral genomes
- Found many NEW species
- Made a catalogue of viral protein clusters and annotated using HMMER (HMM searches against protein family databases)

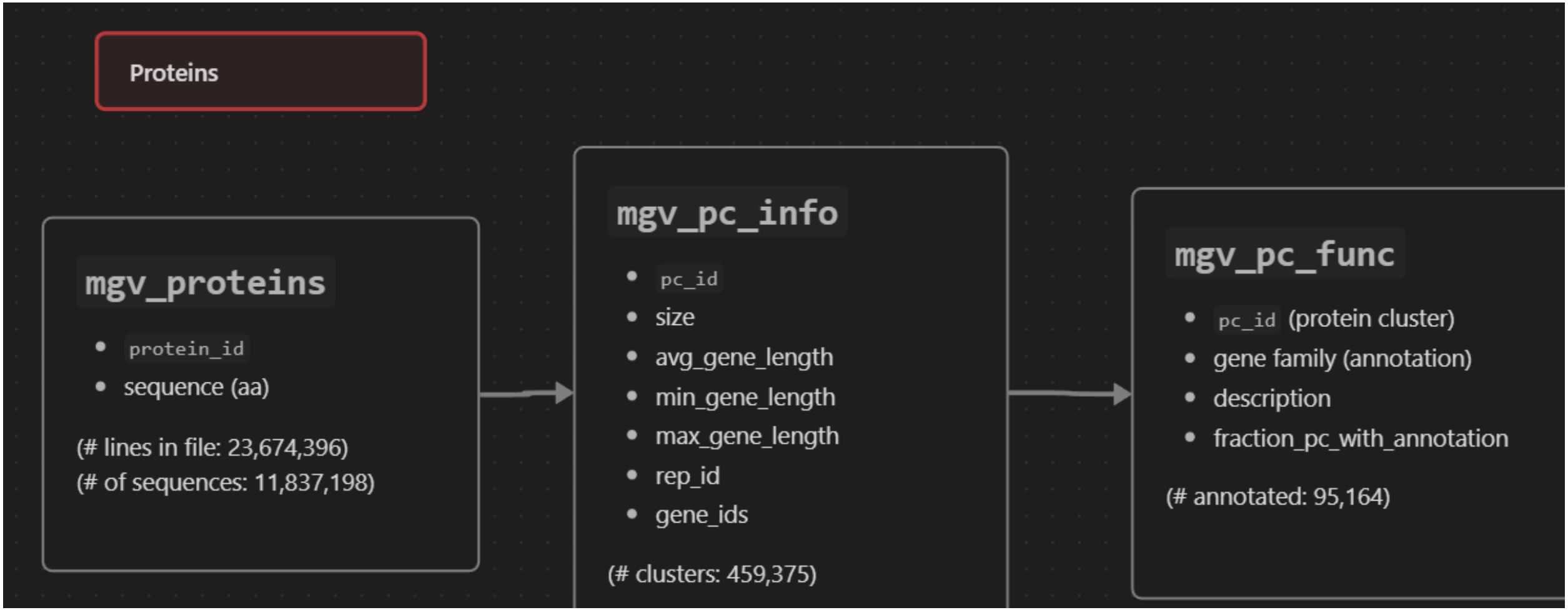
Metadata

mgv_contig_info.tsv

- Metadata for the 189,680 viral genomes. Fields include:
- votu_id: indicate the species-level viral OTU the genome belongs to
- checkv_quality: medium quality (50-90% complete), high quality (>90% complete), complete (closed genome)
- prophage: wheter or not the contig was flanked by DNA from the host (these regions were removed)
- temperate_score: BACPHLIP output indicating the probability the virus lives a temperate lifestyle
- virulent_score: BACPHLIP output indicating the probability the virus lives a virulent lifestyle
- completeness: CheckV estimated completeness
- gc: GC content
- stop_codon_readthrough: indicates whether the virus is predicted to read through a particular stop codon
- baltimore: baltimore classification
- ictv_order, ictv_family, ictv_genus: annotations based on the ICTV taxonomy

contig_id	votu_id	length	checkv_quality	prophage	temperate_score	virulent_score	completeness	gc	stop_codon_readthrough	baltimore	ictv_order	ictv_family	ictv_genus
MGV-GENOME-0364295	OTU-61123	97376	Complete	No	0.0375	0.9625	98.26	31.6166	TAG	dsDNA	Caudovirales	crAss-phage	NULL
MGV-GENOME-0364296	OTU-61123	97376	Complete	No	0.0375	0.9625	98.26	31.6146	TAG	dsDNA	Caudovirales	crAss-phage	NULL
MGV-GENOME-0364303	OTU-05782	97388	Complete	No	0.0357402	0.96426	98.28	27.9706	NULL	dsDNA	Caudovirales	crAss-phage	NULL
MGV-GENOME-0364311	OTU-01114	97394	Complete	No	0.0375	0.9625	98.38	31.4485	TAG	dsDNA	Caudovirales	crAss-phage	NULL
MGV-GENOME-0364312	OTU-23935	97395	Complete	No	0.0138753	0.986125	99.25	33.5777	TAG	dsDNA	Caudovirales	crAss-phage	NULL

Data


Proteins
mgv_proteins
protein_id
sequence (aa)
(# lines in file: 23,674,396)
(# of sequences: 11,837,198)
mgv_pc_info
pc_id
size
avg_gene_length
min_gene_length
max_gene_length
rep_id
gene_ids
(# clusters: 459,375)
mgv_pc_func
pc_id (protein cluster)
gene family (annotation)
description
fraction_pc_with_annotation
(# annotated: 95,164)

Data

- Baltimore Classification:
- 7 classes based on **nucleic acid** (DNA / RNA), **strandness** (double / single), **sense**, **method of replication**

mgv_sample_info

- contig_id
- assembly_source
- assembly_name
- study_accession (# unique: 179,323)
- sample_accession (# unique: 188,684)
- run_accessions
- continent
- country_code
- sex
- age
- health
- disease

mgv_contig_info.tsv.gz

•Metadata for the 189,680 viral genomes. Fields include:

• `contig_id`: indicate the species-level viral OTU the genome belongs to

•checkv_quality: medium quality (50-90% complete), high quality (>90% complete), complete (closed genome)

•prophage: wheter or not the contig was flanked by DNA from the host (these regions were removed)

•temperate_score: BACPHLIP output indicating the probability the virus lives a temperate lifestyle

•virulent_score: BACPHLIP output indicating the probability the virus lives a virulent lifestyle

•completeness: CheckV estimated completeness

•gc: GC content

•stop_codon_readthrough: indicates whether the virus is predicted to read through a particular stop codon

•baltimore: baltimore classification

•ictv_order, ictv_family, ictv_genus: annotations based on the ICTV taxonomy

mgv_contigs

- `contig_id`
- sequence (DNA)

(# lines: 189,681)
(# sequences: 189,680)

mgv_host_assignments.tsv.gz

- contig
- host: (# unique: 246)
- host_phylum: (# unique: 102000)
- host_class: (# unique: 102197)
- host_order: (# unique: 127,548)
- host_family: (# unique: 145047)
- host_genus: (# unique: 141839)
- host_species: (# unique: 112148)

(# lines: 170,093)

mgv_votu_representatives

- contig_id
- vOTU

(# vOTUs: 54,118)

Annotations in Dataset

- **Paper**:
  - Clustered data using MMseq2
  - Annotated 20% using HMMER:
    - HMMER: detects homology by comparing a profile-HMM (a Hidden Markov model constructed explicitly for a particular search) to either a single sequence or a database of sequences.
- **Pfam**:
  - collection of protein families (MSA and HMMs)
- **Results**
  - 411 portal proteins
  - 146 connector

annotation	vpc_id	protein_id	protein_seq
Phage gp6-like head-tail connector protein	VPC-8627	MGV-GENOME-0282701_34	MSLDDEKILEKIKFSCRIDDDI
Phage gp6-like head-tail connector protein	VPC-16699	MGV-GENOME-0270537_10	MLSMADFEDTVLINVKEDLA
Phage gp6-like head-tail connector protein	VPC-135993	MGV-GENOME-0232097_34	MSIKNLMGTVTDDDLQLTKT
Phage gp6-like head-tail connector protein	VPC-545	MGV-GENOME-0260596_65	MEYTTLEQVKIRLKQFHIDTV
Phage gp6-like head-tail connector protein	VPC-456140	MGV-GENOME-0209946_11	MSGEAAAFKPPNRTERTKER

Workflow

1. **Find connector protein domain (in literature / NCBI)**
2. **BLAST**
3. **Filter alignment results by e-value (want very low e-values and high bit scores)**
   - Have to make decisions based on the data
4. **Check if the proteins are annotated / clustered (mgv_pc_info.tsv.gz)**
5. **Model some proteins in list / cluster (to make sure the results are correct)**

1. Connector Domain

- Used Phage connector domain (from NCBI)
- Did BLAST: ~520 sequences
- Filtered by e-value: 2e-60 as a threshold (was just the highest “significant” e-value)
- Found protein cluster (in database) that matched most of the BLAST outputs (5464 sequences)

Sequence Alignment — include consensus sequence ?

Reformat | Format: Hypertext | Row Display: All 4 rows | Color Bits: 2.0 bit | Type Selection: top listed sequences

```
1IJG_I   8 TYRS----INEIQRQK----RNR--WFIHYLNYLQSLAYQLFEWENLPPTINPSFLEKSIHQFGYVGFYKDPVISYIACN 77  Bacillus virus phi29
Q37891   7 SYKS----INDIQRMR----GNR--WYYHYYQYLCSLAYQLFEWERLPPSVDPSYLEKSIHQFGYVGFYKDPRIGYIACQ 76  Bacillus virus B103
Q37995   2 SYKNykrhLGKIELNKetveRNRlaFFEFYFNYFYNIVVNYFTWEGLPNDIDELFIEKKLIENGHVAFFHDDTFGYIAQG 81  Streptococcus phage Cp-1
Q9FZW5   7 SYKT----IGEIQRRR----GNL--WFRTYQRYLFSLAYQMFEWQGLPKTVDPIFLEKQLHQRGFVAFYKDEMYGYLGVQ 76  Bacillus virus GA1

1IJG_I  78 GALSGQRDVYNQATVFR---AASPVYQKEFKLYN----YR---DMKEEDMG------VVIYNNDMAFPTTPTLELFAAEL 141 Bacillus virus phi29
Q37891  77 GALSGTVDHYNLPDRFH---ASSVGYQNTFKLYN----YS---DMKEKNMG------VAIYNNDLKCSTLPALEMFAQDL 140 Bacillus virus B103
Q37995  82 GTRGERLNHYDQPLTYQpvnASSMNYFKQMEIAYtendFRvieELHKDNPDkikrpcIVIPNNNFYEPYIGYLELFCEKL 161 Streptococcus phage Cp-1
Q9FZW5  77 GTLSGQINLYNQPNFYT---ASAPTYQKSFPLYW----YDmgeDLNEKGQG------IVIYNNLERMPTLDILNLYAMNL 143 Bacillus virus GA1

1IJG_I 142 AELKEIISVNQNAQKTPVLIRANDNNQLSLKQVYNQYEGNAPVIFAHEALD----------SDSIEVFKTDAPYVVDKLN 211 Bacillus virus phi29
Q37891 141 AELKEIIAVNQNAQKTPVLIAANDNNQLSLKNIYNQYEGNAPVIFVHESLD----------LDNLKVFKTDAPYVVDKLN 210 Bacillus virus B103
Q37995 162 ADIELTIQLNRNAQITPYFIFADNTNVLSMKNIFNKIANFEPVVYLNKQKDqdgqdsfkqlSDYIQVFRTDAPFLLDKLH 241 Streptococcus phage Cp-1
Q9FZW5 144 AELKETIYVNQNAQKTPVIIKAGDNDLFSMKQVYNKYEGNEPVIFAGKKFN----------TDDIEVLKTDAPYVADKLT 213 Bacillus virus GA1

1IJG_I 212 AQKNAVWNEMMTFLGIKNANLEKKERMVTDEVSSNDEQIESSGTVFLKSREEACEKINELYGLNVKVKFRYDIV 285 Bacillus virus phi29
Q37891 211 AQKNAVWNEVMTYLGIKNANLEKKERMVTSEVDSNDEQIESSGNIYLKARQEACNKISELYGLNLKVKFRYDIV 284 Bacillus virus B103
Q37995 242 DEKLRVMNQLLTFIGINNNPSDKKERLVVSEAISNNGVISANIEVGWKSRRKFVELINKCYGLEISVKPAETIQ 315 Streptococcus phage Cp-1
Q9FZW5 214 MLFKDQWNEAMTFLGLSNANTDKKERLIQSEVESNNDQIQGSANIYLAPRQEACRLINEYYGLNVSVKLRKELV 287 Bacillus virus GA1
```

2&3. BLAST

Sequences producing significant alignments:	Score (Bits)	E Value
MGV-GENOME-0212193_24 # 19174 # 20226 # 1 # ID=267_24;partial=00;...	140	2e-36
MGV-GENOME-0159433_29 # 20711 # 21538 # 1 # ID=491_29;partial=00;...	138	2e-36
MGV-GENOME-0210500_19 # 15471 # 16532 # 1 # ID=861_19;partial=00;...	139	3e-36
MGV-GENOME-0117354_6 # 5749 # 6810 # 1 # ID=496_6;partial=00;star...	139	4e-36
MGV-GENOME-0222640_14 # 9874 # 10935 # -1 # ID=1055_14;partial=00...	136	4e-35
MGV-GENOME-0209211_21 # 16647 # 17711 # 1 # ID=1419_21;partial=00...	135	1e-34
MGV-GENOME-0131812_19 # 12507 # 12902 # -1 # ID=367_19;partial=00...	76.6	9e-15
MGV-GENOME-0191353_6 # 4015 # 4557 # -1 # ID=2682_6;partial=00;st...	75.1	7e-14
MGV-GENOME-0191353_5 # 3534 # 3938 # -1 # ID=2682_5;partial=00;st...	63.9	3e-10
MGV-GENOME-0105632_1 # 3 # 485 # -1 # ID=725_1;partial=10;start_t...	58.9	3e-08
MGV-GENOME-0214625_28 # 24444 # 25526 # 1 # ID=514_28;partial=00;...	55.5	3e-06
MGV-GENOME-4395318_4 # 2697 # 3779 # 1 # ID=1794_4;partial=00;sta...	53.5	1e-05
MGV-GENOME-0215696_35 # 28616 # 29707 # -1 # ID=1451_35;partial=0...	51.6	5e-05

- Here, I chose a point when the Bit score drops off
- After checking if the results correspond to any protein clusters, I found VPC_8016, which has exactly the same number of proteins in the cluster as there are in the BLAST results (269 proteins)
- E-value = number of expected hits of similar quality (score) that could be found just by chance
- Bit score = size of database you would need to see an alignment by chance
- You want smaller e-values and larger bit scores

4. Check for annotation

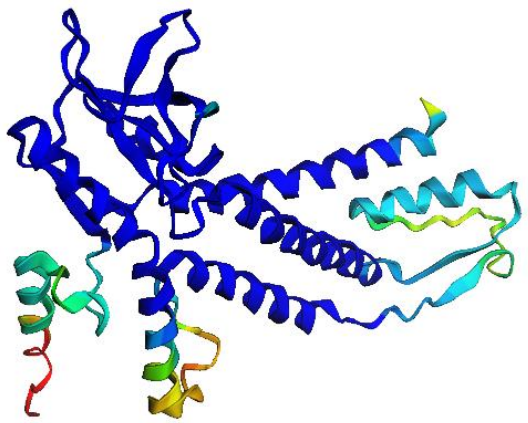
- The BLAST results (using the connector domain as a query) match the RNA ligase annotation for some reason

```
cat filtered_connector_domain.txt | cut -f 1 -d " " | while read line; do
    vpc=$(zgrep -m 1 $line mgv_pc_info.tsv.gz  | cut -f 1)
    function=$(zgrep -m 1 $vpc mgv_pc_functions.tsv.gz | cut -f 3)
    echo "$line $vpc $function"
done
```

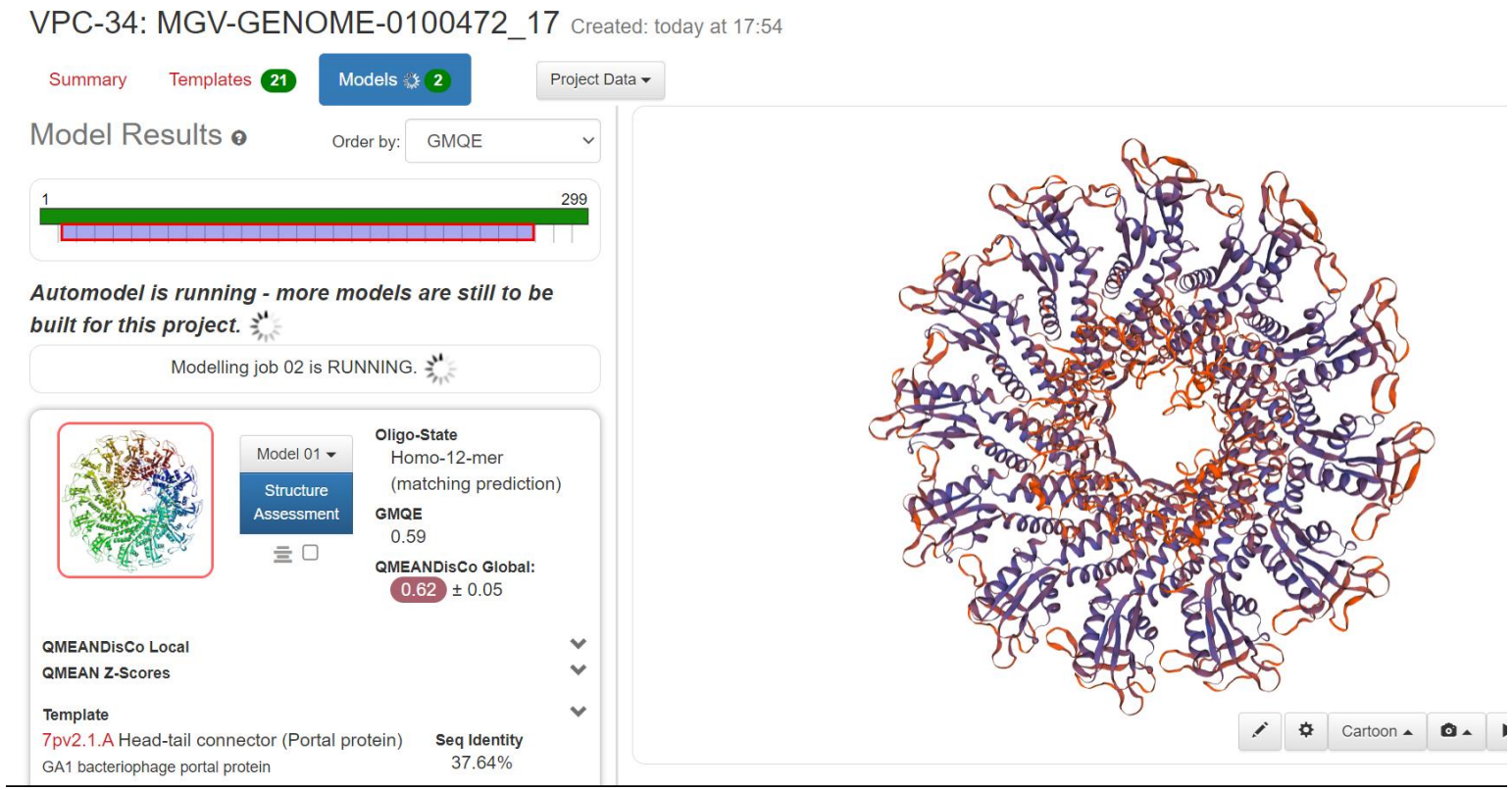
```
MGV-GENOME-0104393_9 VPC-34  RNA ligase
MGV-GENOME-0122635_24 VPC-34  RNA ligase
MGV-GENOME-4313378_4 VPC-34  RNA ligase
MGV-GENOME-0094600_4 VPC-34  RNA ligase
MGV-GENOME-0094502_14 VPC-34  RNA ligase
MGV-GENOME-0095706_5 VPC-34  RNA ligase
MGV-GENOME-0099638_13 VPC-34  RNA ligase
MGV-GENOME-0103984_11 VPC-34  RNA ligase
MGV-GENOME-0118545_23 VPC-34  RNA ligase
MGV-GENOME-0081748_6 VPC-34  RNA ligase
MGV-GENOME-0080382_13 VPC-34  RNA ligase
MGV-GENOME-0087131_3 VPC-34  RNA ligase
MGV-GENOME-0103282_15 VPC-34  RNA ligase
MGV-GENOME-0125157_6 VPC-34  RNA ligase
MGV-GENOME-0052998_5 VPC-34  RNA ligase
```

5. Modelling

- **Modelled some of the proteins found in the protein cluster (VPC-34)**
- **Result:**
- **AlphaFold:**



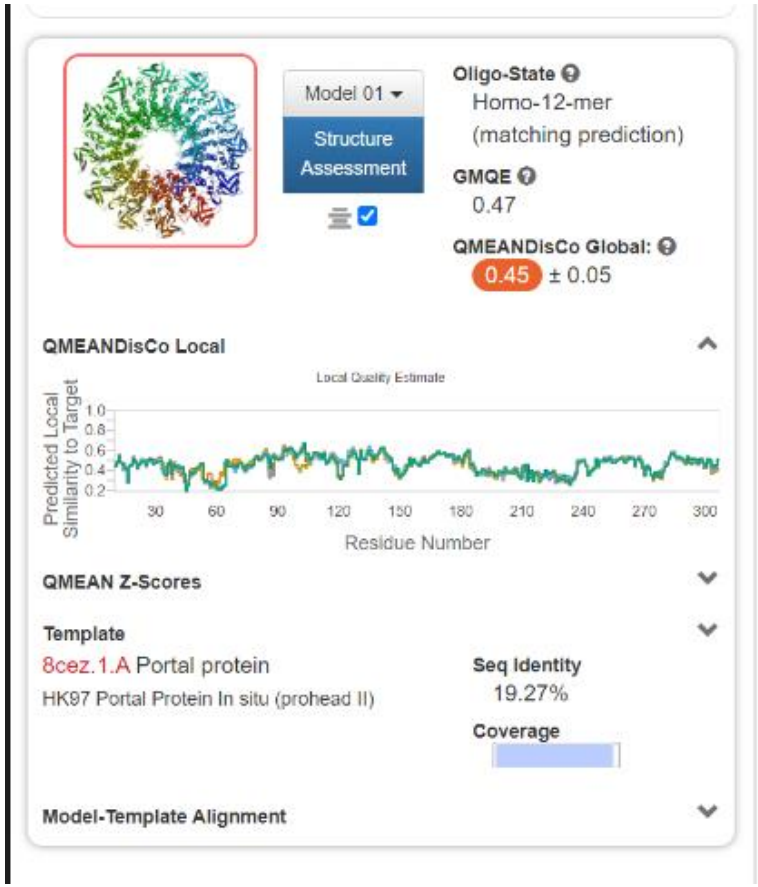
- Swiss-MODEL:

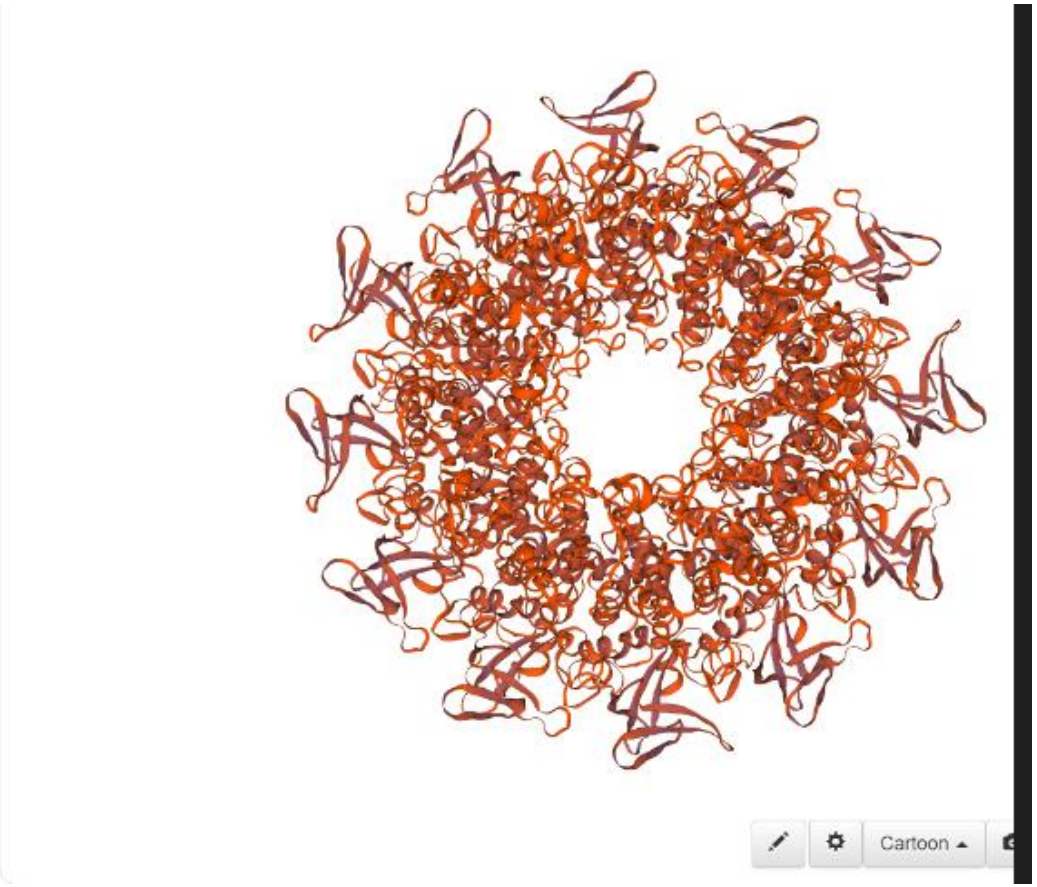


pIDDT: Very low (<50) Low (60) OK (70) Confident (80) Very high (>90)

Query Sequence

>YP_010091615.1 portal protein [uncultured Caudovirales phage]





BLAST: Query Sequence

```
(base) claireh@claire-virtualbox:/media/sf_shared_folder/Bacteriophage/MGV$ cat blast_protein/filtered_query.txt
MGV-GENOME-0192513_23 # 17497 # 18534 # 1 # ID=3086_23;partial=00...  172     8e-49
MGV-GENOME-0212292_17 # 12116 # 13165 # 1 # ID=320_17;partial=00;...  170     5e-48
MGV-GENOME-0172516_15 # 10958 # 12007 # -1 # ID=1136_15;partial=0...  170     5e-48
MGV-GENOME-0210072_17 # 11748 # 12797 # 1 # ID=55_17;partial=00;s...  170     5e-48
MGV-GENOME-0226859_40 # 31378 # 32427 # 1 # ID=795_40;partial=00;...  170     6e-48
MGV-GENOME-0187969_20 # 13343 # 14392 # -1 # ID=372_20;partial=00...  169     1e-47
MGV-GENOME-0183924_16 # 11035 # 12084 # -1 # ID=1677_16;partial=0...  169     2e-47
```

```
(base) claireh@claire-virtualbox:/media/sf_shared_folder/Bacteriophage/MGV$ cat blast_protein/filtered_connector_dom
MGV-GENOME-0104393_9 # 6790 # 7770 # 1 # ID=513_9;partial=00;star...  254     6e-81
MGV-GENOME-0122635_24 # 15385 # 16266 # 1 # ID=1448_24;partial=00...  236     1e-74
MGV-GENOME-4313378_4 # 3458 # 4342 # -1 # ID=783_4;partial=00;sta...  233     3e-73
MGV-GENOME-0094600_4 # 3582 # 4466 # -1 # ID=892_4;partial=00;sta...  233     3e-73
MGV-GENOME-0094502_14 # 10883 # 11767 # 1 # ID=877_14;partial=00;...  233     3e-73
MGV-GENOME-0095706_5 # 3584 # 4468 # -1 # ID=1053_5;partial=00;st...  231     2e-72
MGV-GENOME-0099638_13 # 9158 # 10123 # 1 # ID=1418_13;partial=00;...  231     2e-72
```

- BLAST results using query sequence aren’t quite as good as using connector domain but they’re still good

Query Seq. (cont)

Annotation

```
MGV-GENOME-0192513_23 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0212292_17 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
SMGV-GENOME-0172516_15 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0210072_17 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0226859_40 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0187969_20 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0183924_16 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0165836_19 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
MGV-GENOME-0142829_6 VPC-8016  Cytidine and deoxycytidylate deaminase zinc-binding region
```

Query Seq. (cont)

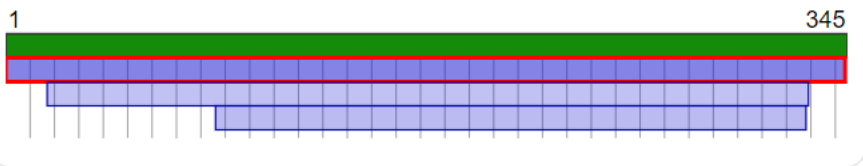
MGV-GENOME-0192513_23, labelled as Cytidine and deoxycytidylate deaminase zinc-binding region

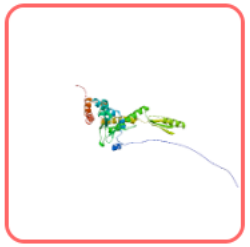
Created: Nov. 21, 2023 at 03:29

Summary | Templates 20 | Models 3 | Project Data

Model Results

Order by: GMQE





Model 01

Structure Assessment

Oligo-State
Monomer

GMQE
0.80

Template

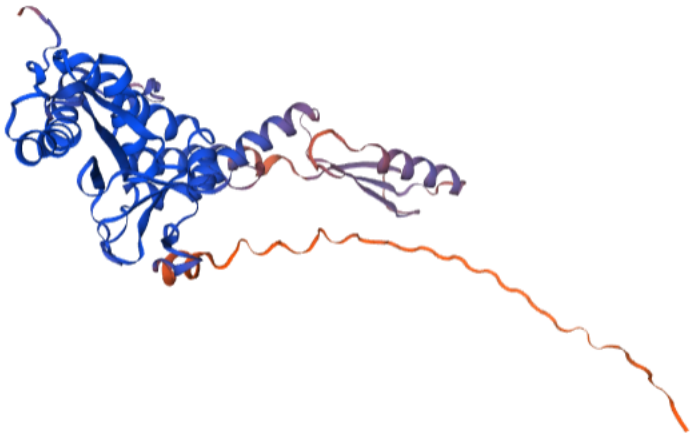
A0A3Z9LJW2.1.A Phage portal protein

AlphaFold DB model of A0A3Z9LJW2_SALER (gene: A0A3Z9LJW2_SALER, organism: Salmonella enterica (Salmonella choleraesuis))

Seq Identity
79.94%

Coverage

Model-Template Alignment



Cartoon

Ideas for how to find the rest of the connector sequences

- ML Programs for Protein Annotation:
  - http://phanns.com/
    - Doesn’t work? Problem with FASTA headers
  - http://prodata.swmed.edu/MESSA/MESSA.cgi
    - Need query organism?
- Looking at protein structure