BACTERIOPHAGE GENOMICS: A TOOL FOR LEARNING BIOINFORMATICS

JOSEPH CHRISTIANSEN, BETH WILKES
DEPARTMENT OF NATURAL SCIENCES

Background

Bacteriophages are viruses that infect bacteria. Image 1 depicts a phage's morphology. They are a potential alternative to antibiotics in fighting infections, and their small genome lends themselves well to studying bioinformatics [1]. The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program provides undergraduates an opportunity for research-based coursework involving bacteriophages [2]. The students isolate a phage and annotate its genome. The purpose of this project is to outline the process of phage genome annotation; gene 24 from "Damp," a phage isolated by the University of Pittsburgh, depicts the process. Damp is a *Gordonia* phage (i.e., it infects *Gordonia* bacteria). Genome annotation can serve as an introduction to bioinformatics. Bioinformatics involves storing and interpreting biological data (e.g., nucleotide sequences), which can be used to make functional predictions.

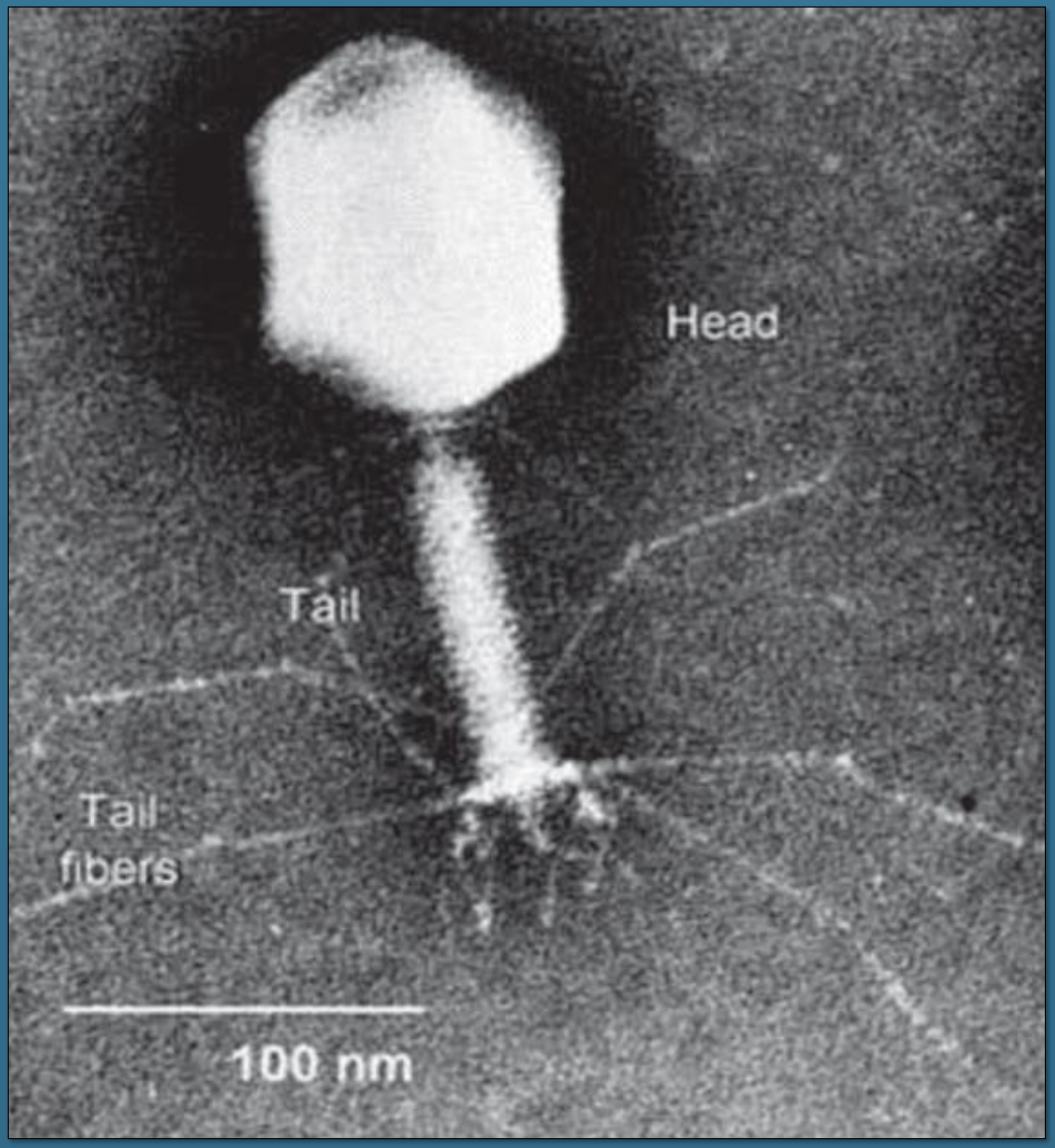


Image 1 – TEM of a bacteriophage. The head contains the DNA, and the tail fibers help bind to the bacteria. Photo: [3]

Manual Annotation

Auto-annotation programs will misinterpret or fail to call 5-10 genes per bacteriophage genome [4]. Because of this, it is paramount that a review of the auto-annotation (i.e., the manual annotation) occurs.

Part 1: Gene Start Prediction

The manual annotation analyzes outputs from various algorithms to confirm or edit the auto-annotation's chosen gene start. Coding potential, gaps, and ribosomal binding site (RBS) scores are assessed.

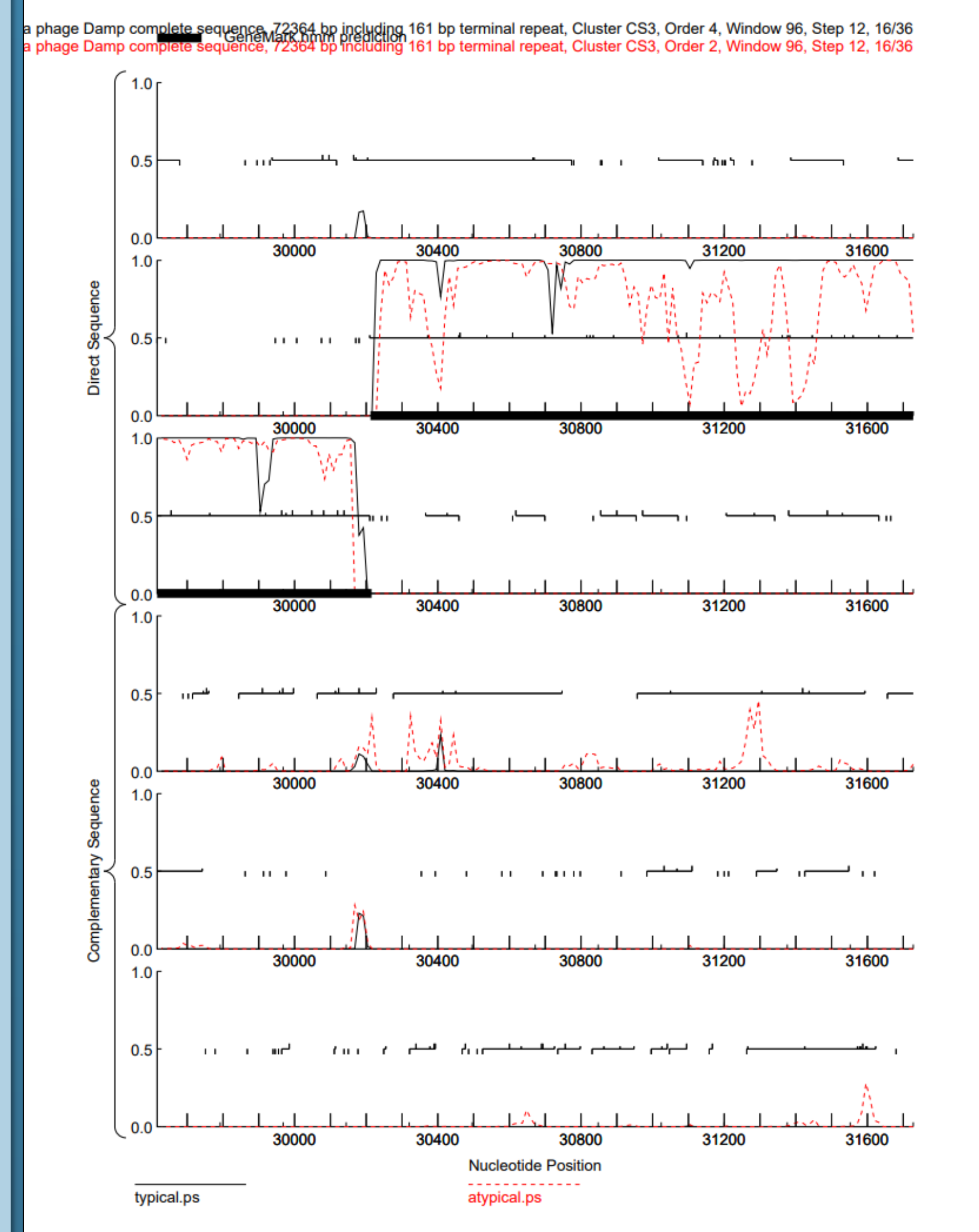
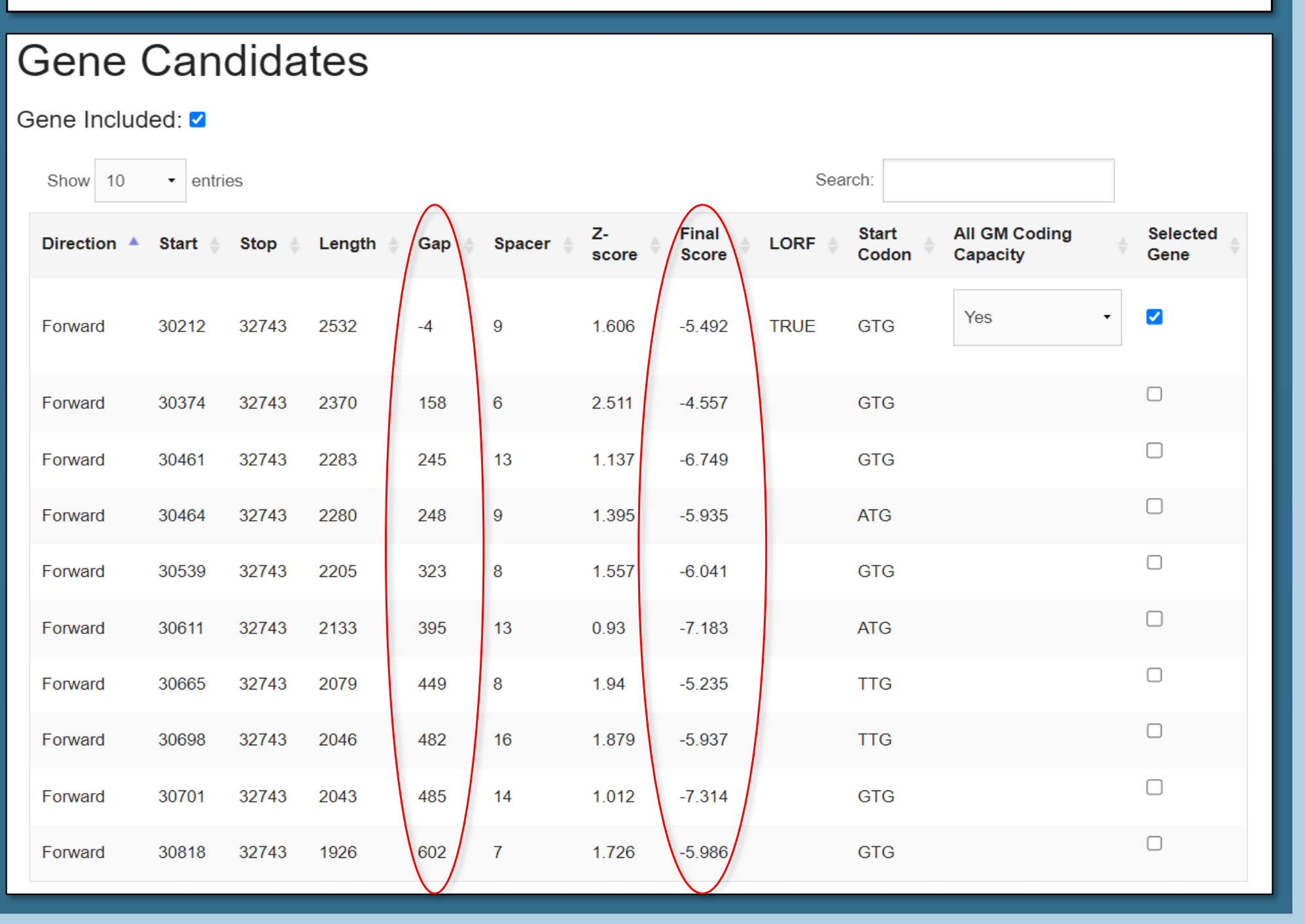


Image 2 (left) – GeneMarkS [5] graph of gene 24 for Damp (Damp_24). The y-axis represents coding potential, the probability that a gene lies at a locus. The x-axis represents the nucleotide position. Gene starts that capture all the coding potential are likely to be the best start for a gene. The graph displays 6 reading frames. A reading frame is the pattern of codons (i.e., triplets of nucleotides) used during translation. The top 3 reading frames are in the forward direction, and the bottom 3 are in the reverse. Red dashed lines represent predictions with an atypical model, and solid lines represent predictions using a typical model; coding potential with the typical model is more robust evidence.

Image 3 (right) – Table of gene starts and their gaps and RBS scores (final score column) [6]. A low gap is stronger evidence for a gene start, as phages favor a compact genome. An RBS score relays the strength of a ribosomal binding site for a gene start. Ribosomal binding sites are sequences of DNA responsible for ensuring translational accuracy. A high score is strong evidence for a gene start.



Gene Candidates

Gene Included: ☑

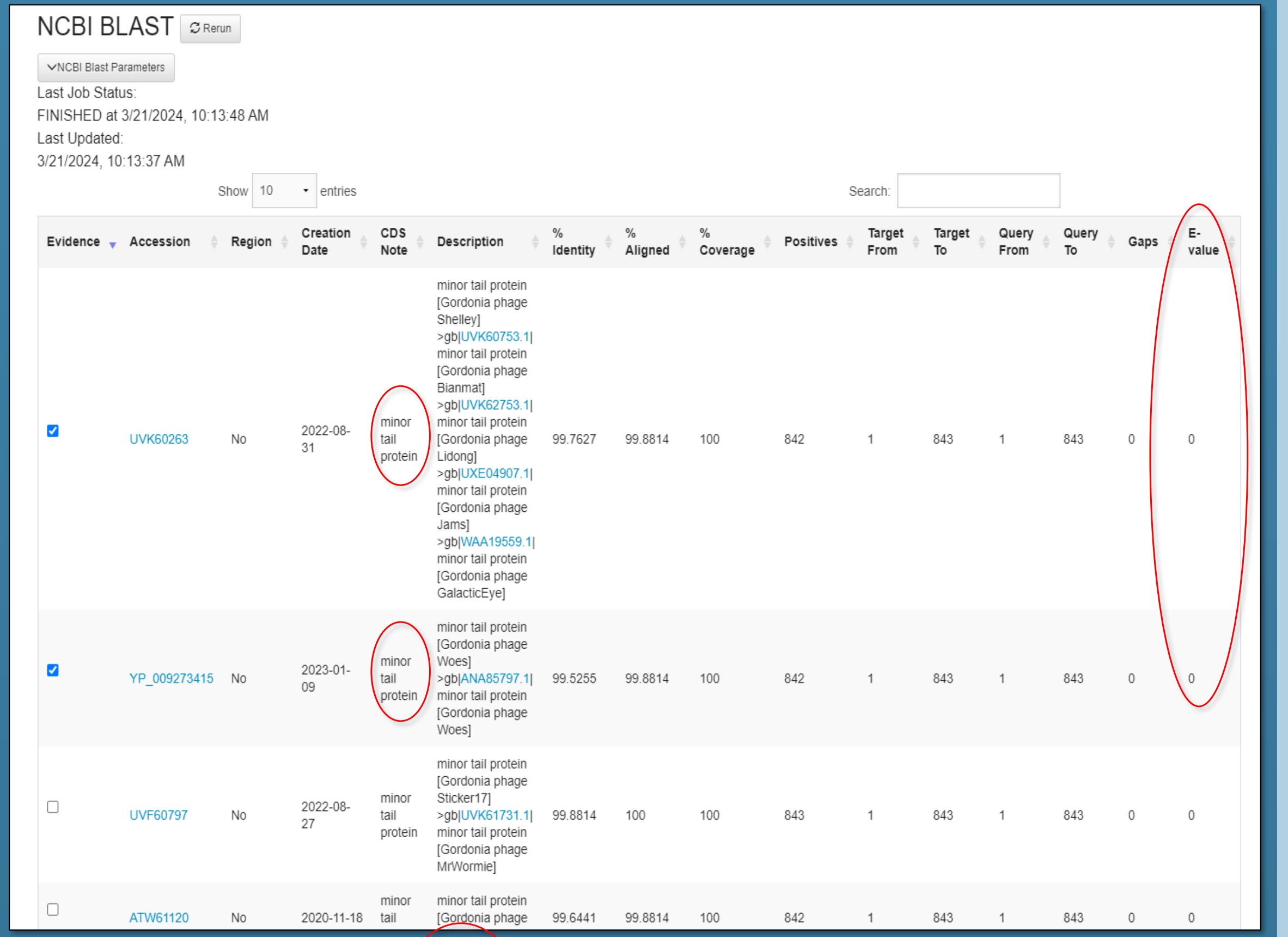
Show 10 entries   Search:

Direction	Start	Stop	Length	Gap	Spacer	Z-score	Final Score	LORF	Start Codon	All GM Coding Capacity	Selected Gene
Forward	30212	32743	2532	-4	9	1.606	-5.492	TRUE	GTG	Yes	☑
Forward	30374	32743	2370	158	6	2.511	-4.557		GTG		□
Forward	30461	32743	2283	245	13	1.137	-6.749		GTG		□
Forward	30464	32743	2280	248	9	1.395	-5.935		ATG		□
Forward	30539	32743	2205	323	8	1.557	-6.041		GTG		□
Forward	30611	32743	2133	395	13	0.93	-7.183		ATG		□
Forward	30665	32743	2079	449	8	1.94	-5.235		TTG		□
Forward	30698	32743	2046	482	16	1.879	-5.937		TTG		□
Forward	30701	32743	2043	485	14	1.012	-7.314		GTG		□
Forward	30818	32743	1926	602	7	1.726	-5.986		GTG		□

Part 2: Protein Function Prediction

After predicting where a gene starts, the gene product (i.e., the protein) can be assessed. This involves matching nucleotide or amino acid sequences against databases.

Image 4 (right) – The NCBI basic local alignment search tool (BLAST) [7] matches a gene's nucleotide sequence against the NCBI database. It may target an entire gene or a portion of one. Matches with known gene products indicate possible protein function. The e-value for each hit represents the probability that an alignment occurred by chance. Therefore, hits with lower e-values are stronger evidence for function.

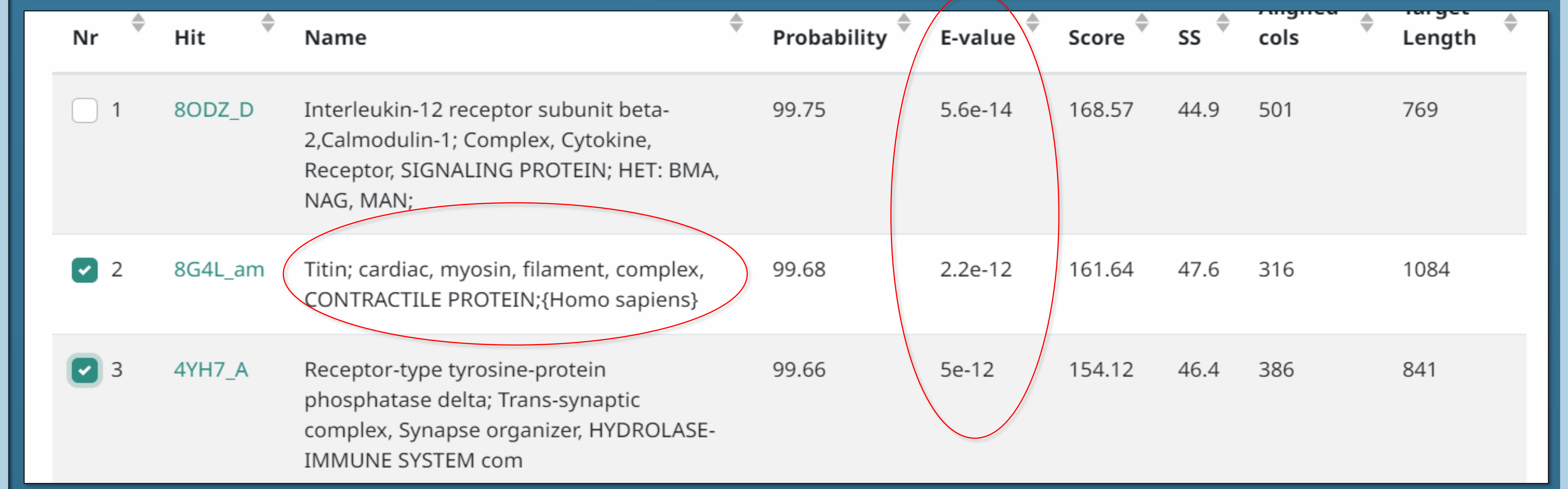


NCBI BLAST

Last Job Status:
FINISHED at 3/21/2024, 10:13:48 AM
Last Updated:
3/21/2024, 10:13:37 AM

Show 10 entries   Search:

Evidence	Accession	Region	Creation Date	CDS Note	Description	% Identity	% Aligned	% Coverage	Positives	Target From	Target To	Query From	Query To	Gaps	E-value
☑	UVK60263	No	2022-08-31	minor tail protein	minor tail protein [Gordonia phage Shelley] >gb|UVK60753.1| minor tail protein [Gordonia phage Bianmat] >gb|UVK62753.1| minor tail protein [Gordonia phage Lidong] >gb|UXE04907.1| minor tail protein [Gordonia phage Jams] >gb|WAA19559.1| minor tail protein [Gordonia phage GalacticEye]	99.7627	99.8814	100	842	1	843	1	843	0	0
☑	YP_009273415	No	2023-01-09	minor tail protein	minor tail protein [Gordonia phage Woes] >gb|ANA85797.1| minor tail protein [Gordonia phage Woes]	99.5255	99.8814	100	842	1	843	1	843	0	0
□	UVF60797	No	2022-08-27	minor tail protein	minor tail protein [Gordonia phage Sticker17] >gb|UVK61731.1| minor tail protein [Gordonia phage MrWormie]	99.8814	100	100	843	1	843	1	843	0	0
□	ATW61120	No	2020-11-18	minor tail	minor tail protein [Gordonia phage	99.6441	99.8814	100	842	1	843	1	843	0	0



Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
□ 1	8ODZ_D	Interleukin-12 receptor subunit beta-2,Calmodulin-1; Complex, Cytokine, Receptor, SIGNALING PROTEIN; HET: BMA, NAG, MAN;	99.75	5.6e-14	168.57	44.9	501	769
☑ 2	8G4L_am	Titin; cardiac, myosin, filament, complex, CONTRACTILE PROTEIN;{Homo sapiens}	99.68	2.2e-12	161.64	47.6	316	1084
☑ 3	4YH7_A	Receptor-type tyrosine-protein phosphatase delta; Trans-synaptic complex, Synapse organizer, HYDROLASE-IMMUNE SYSTEM com	99.66	5e-12	154.12	46.4	386	841

Image 5 – HHpred [8] matches amino acid sequences against protein databanks. It uses a different algorithm than NCBI BLAST. HHpred is important to protein calls, as it may align to wet lab results, which is strong evidence. HHpred also provides the e-value for a particular hit. This query for Damp_24 is matching to a contractile protein, which is evidence for a minor tail functional call.

References

[1] Abedon, S. T., Kuhl, S. J., Blasdel, B. G., & Kutter, E. M. (2011). Phage treatment of human infections. *Bacteriophage*, *1*(2), 66–85. https://doi.org/10.4161/bact.1.2.15845
[2] SEA-PHAGES. (n.d.). https://seaphages.org/
[3] Abdul Wahid, A. (2015). Phage Therapy: Emergence of a Novel Therapy to Control Bacterial Pathogens. *Inventi Rapid: Pharm Biotech & Microbio*, *1*, 1.
[4] Poxleitner, M., Pope, W., Jacobs-Sera, D., Sivanathan, V., Graham., H. Phage Discovery Guide. Howard Hughes Medical Institute. (2018) https://seaphagesphagediscoveryguide.helpdocsonline.com/home
[5] Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, Vol. 29, No. 12, 2607-2618
[6] Claire A. Rinehart, Bobby Gaffney, James Dexter Wood, Jason Smith. (2016). PECANN, a Phage Evidence Collection and Annotation Network. https://discover.kbrinsgd.org
[7] Camacho C, Boratyn GM, Joukov V, Vera Alvarez R, Madden TL. ElasticBLAST: accelerating sequence search via cloud computing. *BMC Bioinformatics*. 2023 Mar 26;24(1):117. https://doi.org/10.1186/s12859-023-05245-9. PMID: 36967390
[8] Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A. N., & Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics*, *72*(1), e108. https://doi.org/10.1002/cpbi.108

Acknowledgements

Special thanks to:

Professor Beth Wilkes for advice on genome annotation & project feedback.

Professors Tracey Lesser and Jennifer Eggers for project feedback.

Classmates for project feedback.