Discarded diversity: Novel megaphages, auxiliary metabolic genes, and virally encoded CRISPR-Cas systems in landfills

Nikhil A. George[1], Zhichao Zhou[2], Karthik Anantharaman[2], Laura A. Hug[1,*]

Supplemental Materials

Table of Contents:

Supplemental Results

Similarity to previously identified virally encoded CRISPR-Cas systems

We observed several instances where our predicted effector sequences were clustered with or were most closely related to previously identified virally encoded CRISPR-Cas systems (i.e., Cas14j (7), Cas14k (9), Cas14i (2), Cas12j (1), and Cas12L (1), Figure 4), of which only Cas12j and Cas12L have been experimentally validated for function (Pausch et al. 2020; Al-Shayeb et al. 2022). Each of our relevant sequences was examined for RuvC motifs and assessed for similarity to the virally encoded nuclease it clustered most closely with.

Seven sequences ranging from 373-441aa clustered with previously identified Cas14j sequences (378-451aa; Figure 4, 10 o'clock). Our sequences showed high sequence similarity to

Cas14j sequences and contained all three RuvC motifs (RuvCI-III). Six additional sequences clustered proximal but distinct to the Cas14J cluster and are described in the next section. We detected two sequences of lengths 402 and 509 aa that clustered with Cas14i (Figure 4, 11 o'clock). Both of these sequences as well as reference Cas14i proteins had detectable RuvCI and RuvCIII motifs but very weak, if detectable, RuvCII motifs. We detected one sequence clustering with Cas12j (Figure 4, 7 o'clock). This sequence lacked the RuvCIII motif, as did Cas12j6, one of the 10 reference Cas12j nucleases (Al-Shayeb et al. 2020). Our putative Cas12j sequence was also missing key residues in the RuvCII domain, which Cas12j6 lacked entirely, and key residues in the RuvCI motif, all of which were present in Cas12j6. Only three orthologs of Cas12j(1-3) have been experimentally confirmed for function (Pausch et al. 2020). The comparisons made to Cas12j6 add confidence to the assignment of our query sequence as a Cas12j ortholog, despite its lack of key catalytic residues. Our putative Cas12j is the shortest within the clade, at 346 aa compared to 441 (Cas12j6 from a giant phage (Al-Shayeb et al. 2020)) and 708-813aa for the remaining 9 Cas12j proteins. While our sequence branches within the Cas12j clade, its activity is less confidently predicted based on the aberrant characteristics described above. Notably, our putative Cas12j sequence is encoded by a predicted plasmid, the second time a Cas12j-like protein was identified on a plasmid (Pinilla-Redondo et al. 2022).

Supplemental Tables and Figures

Table S1: Landfill sites and sampling details

Site	Sample ID	Sample type	Metagenome size (Gbp)	BioSample Accession	SRA Accession
SO_2016	LW1	Leachate well	26.58	SAMN07630781	SRX3574636
	LW2	Leachate well	30.00	SAMN07630782	SRX3574178
	LW3	Leachate well	29.98	SAMN07630780	SRX3574180
	CLC1_T1	Composite leachate cistern	29.89	SAMN07630778	SRX3574177
	CLC1_T2	Composite leachate cistern	28.16	SAMN07630777	SRX3575198
	GW1	Groundwater well	25.58	SAMN07630779	SRX3574179
SO_2017	LW1	Leachate well	15.57	SAMN27259107	SRX14723681
	LW2	Leachate well	40.52	SAMN10350574	SRX5256784
	LW3	Leachate well	47.58	SAMN27259106	SRX14723680
	LW4	Leachate well	38.75	SAMN10863920	SRX5344198
	CLC	Composite leachate cistern	38.72	SAMN10350766	SRX5256785
	SWC	Storm water catchment	21.34	SAMN10350495	SRX5256798
	GW1	Groundwater well	51.22	SAMN27259105	SRX14723679
	GW3	Groundwater well	18.76	SAMN10350765	SRX5256783
NEUS	A	Leachate well	53.99	SAMN31696084	SRX18288880
	B	Leachate well	47.93	SAMN31696085	SRX18288881
	C	Leachate well	56.03	SAMN31696086	SRX18288882
	D1	Leachate well	52.88	SAMN31696087	SRX18288883
	D2	Leachate well	48.60	SAMN31696088	SRX18288884
	E	Leachate well	37.64	SAMN31696089	SRX18288885
	F1	Leachate well	57.67	SAMN31696090	SRX18288886
	F2	Leachate well	50.17	SAMN31696091	SRX18288887
	CSWMC	Composite leachate cistern	54.34	SAMN31696092	SRX18288888
CA_2019	LW1	Leachate well	31.10	SAMN39634476	SRX23416964
	CLC	Composite leachate cistern	64.38	SAMN39634477	SRX23416965
	TP_BF	Treatment plant biofilter - planktonic	61.73	SAMN39634478	SRX23416966
	TP_BS	Treatment plant biofilter - solids	58.01	SAMN39634479	SRX23416967

Table S2: Putative cross-phylum host-virus interactions.

Sample set	Putative hosts	Host MAG phylum (GTDB-tk)	Host completion, contamination (%)	# host spacer to viral protospacer matches	Predicted viral element
CA_2019	TPIn_75 TPBF_198	Desulfobacterota Proteobacteria	99.41, 0.00 90.75, 0.63	1 7	vMAG_518
NEUS_2019	STF2_137 STCSWMC_88	Bacteroidota Firmicutes_A	95.56, 3.26 80.02, 4.08	1 2	vMAG_1257
NEUS_2019	STCSWMC_93 STF1_64 STF2_19 STD2_245 STCSWMC_50 STCSWMC_25	Bacteroidota Bacteroidota Bacteroidota Bacteroidota Cloacimonadota Firmicutes_B	94.35, 0.27 95.43, 0.27 95.43, 0.00 88.71, 2.42 98.90, 2.20 90.15, 4.60	15 16 1 5 1 3	vMAG_3146
NEUS_2019	STF2_137 STCSWMC_88	Bacteroidota Firmicutes_A	95.56, 3.26 80.02, 4.08	1 2	vMAG_910
NEUS_2019	STF2_144 STCSWMC_50 STF2_148 STCSWMC_25	Cloacimonadota Cloacimonadota Cloacimonadota Firmicutes_B	100.00, 1.10 98.90, 2.20 95.54, 1.10 90.15, 4.60	1 2 2 2	NODE_3233_length_30680_cov_384.260596_Sandtown_F2||full
SO_2017	LW2_137 LW2_139	Muirbacteria Patescibacteria	93.26, 4.56 70.53, 3.61	2 4	vMAG_2310

Tables S3 and S4 are included as a single .xlsx file "Supplementary File 1.xlsx"

Table S3: Predicted AMGs encoded across all datasets.
Table S4: AMGs encoded by megaphage genomes.

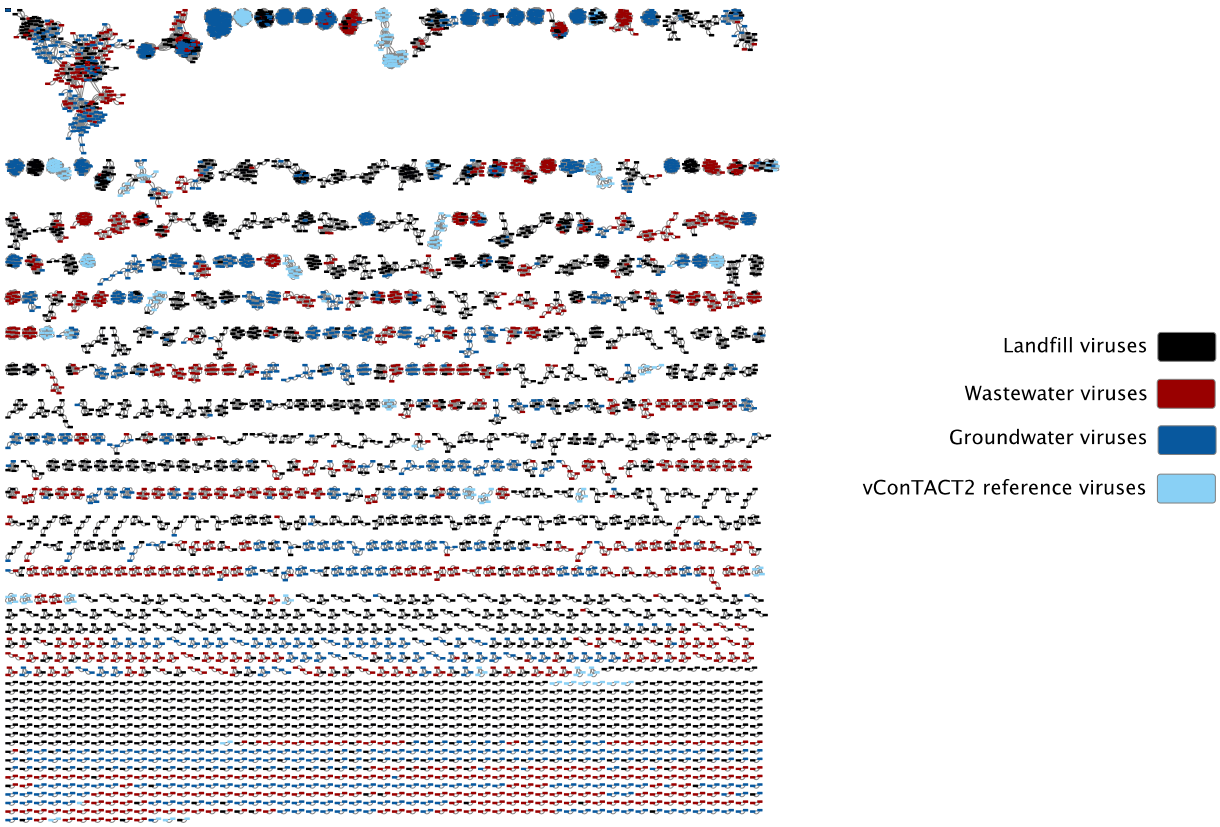

Figure S1: Gene-sharing network of landfill viruses with related groups from IMG/VR. Nodes represent viral elements and are coloured by sample site as summarized in the legend. Nodes connected by edges represent viral elements that share protein clusters. The network was generated using vConTACT2. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy.

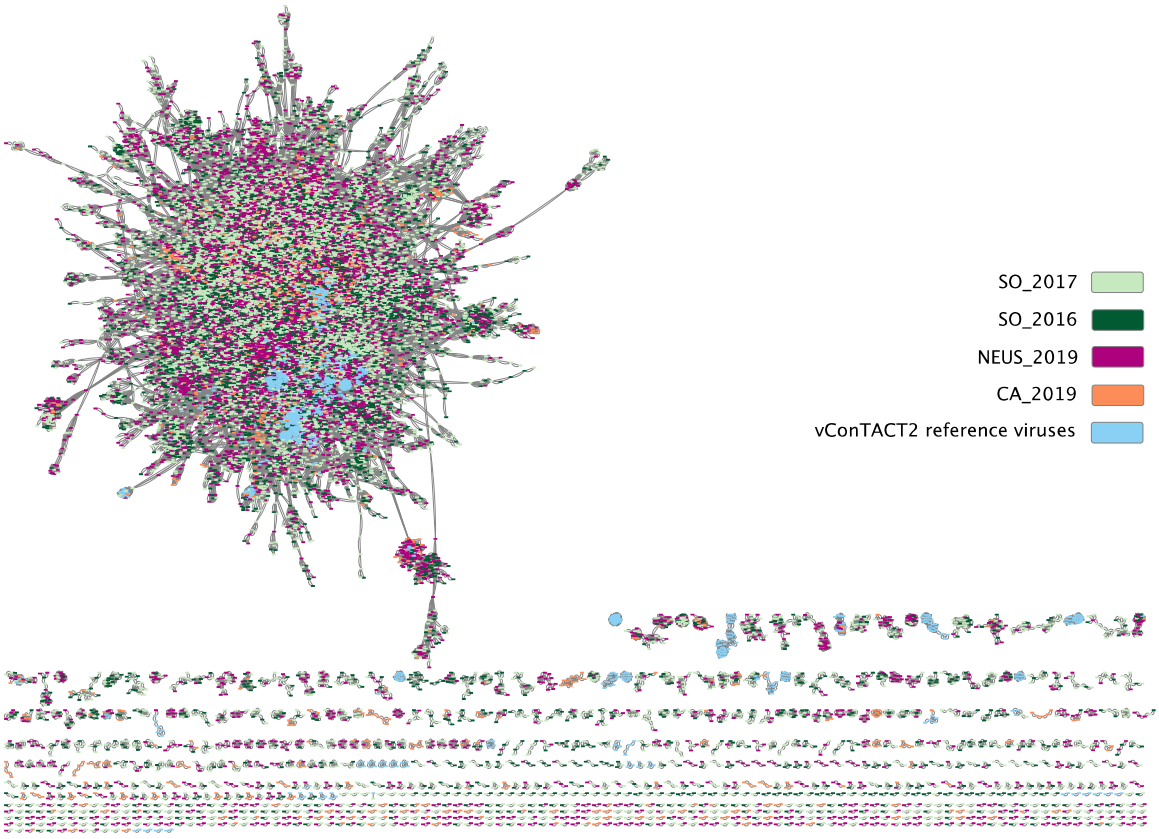

Figure S2: Gene-sharing network of landfill viruses. Viral nodes are coloured by the landfill they were identified in. Nodes connected by edges represent viral elements that share protein clusters. The network was generated using vConTACT2. The vConTACT2 reference database used was Prokaryotic Viral Refseq version 85 with ICTV-only taxonomy.

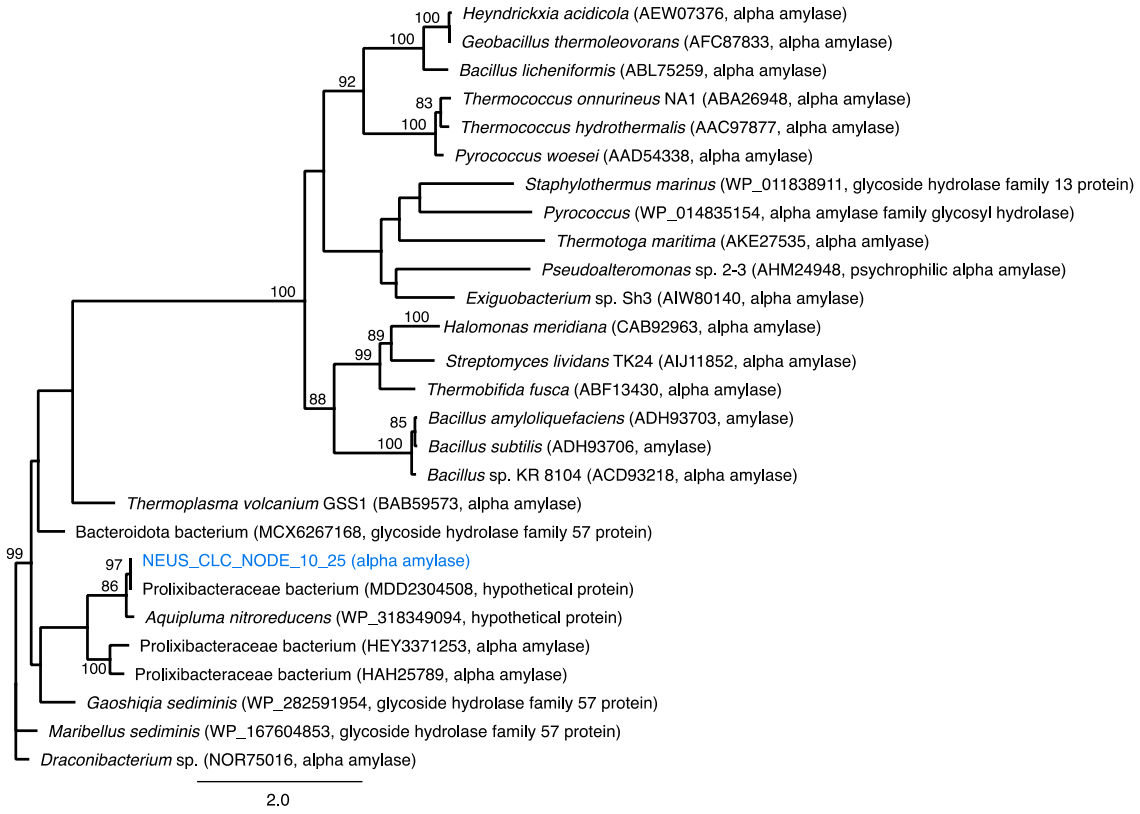

Figure S3: A maximum likelihood phylogeny of the alpha amylase AMG (blue), including representative best hits from a blastp search and reference sequences (Mehta and Satyanarayana 2016). The final alignment contained 28 taxa and 805 unambiguously aligned columns. Alignments were generated with Muscle version 3.8.425 and (Edgar 2004) trimmed to remove columns with more than 90% gaps. The tree was generated using RAxML version 8 under the VT+I+G model of evolution (Stamatakis 2014) and visualized in Geneious (Kearse et al. 2012).

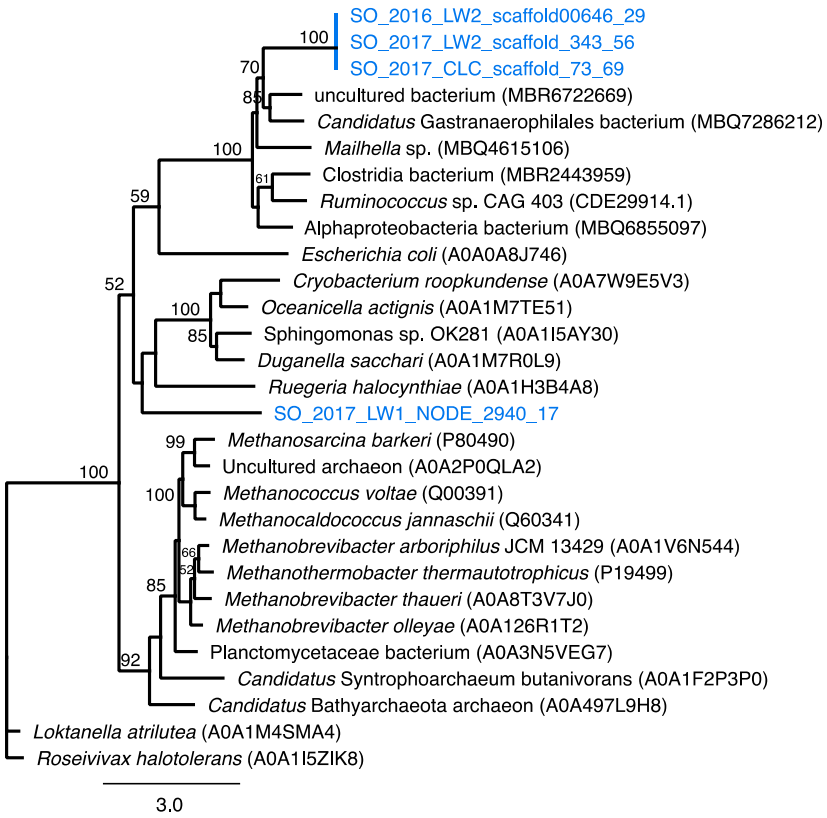

Figure S4: A maximum likelihood phylogeny of the coenzyme P450 hydrolase subunit beta AMGs (blue), including representative best hits from a blastp search and reference sequences from UniProtKB. The final alignment contained 29 taxa and 845 unambiguously aligned columns. Alignments were generated with Muscle version 3.8.425 and (Edgar 2004) trimmed to remove columns with more than 90% gaps. The tree was generated using RAxML version 8 under the LG+I+G model of evolution (Stamatakis 2014) and visualized in Geneious (Kearse et al. 2012).

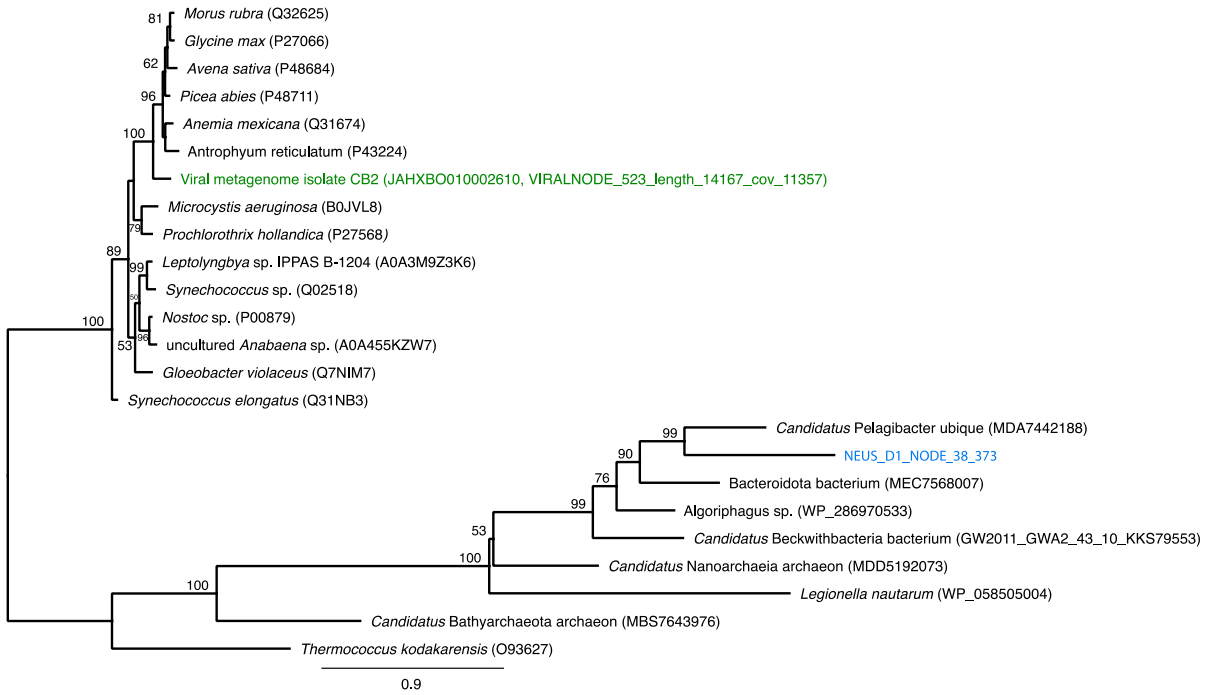

Figure S5: A maximum likelihood phylogeny of the RuBisCO large subunit AMG (blue), including representative best hits from a blastp search and reference sequences from UniProtKB. One previously reported RbcL from a viral fragment was also included (green, Bhattarai et al. 2021). The final alignment contained 24 taxa and 481 unambiguously aligned columns. Alignments were generated with Muscle version 3.8.425 and (Edgar 2004) trimmed to remove columns with more than 90% gaps. The tree was generated using RAxML version 8 under the LG+I+G model of evolution (Stamatakis 2014) and visualized in Geneious (Kearse et al. 2012).

References

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D.S.A., Borton, M.A., Sharrar, A., Jaffe, A.L., Nelson, T.C., Kantor, R., Keren, R., Lane, K.R., Farag, I.F., Lei, S., Finstad, K., Amundson, R., Anantharaman, K., Zhou, J., Probst, A.J., Power, M.E., Tringe, S.G., Li, W.-J., Wrighton, K., Harrison, S., Morowitz, M., Relman, D.A., Doudna, J.A., Lehours, A.-C., Warren, L., Cate, J.H.D., Santini, J.M., and Banfield, J.F. 2020. Clades of huge phages from across Earth's ecosystems. Nature **578**(7795): 425–431. doi:10.1038/s41586-020-2007-4.

Al-Shayeb, B., Skopintsev, P., Soczek, K.M., Stahl, E.C., Li, Z., Groover, E., Smock, D., Eggers, A.R., Pausch, P., Cress, B.F., Huang, C.J., Staskawicz, B., Savage, D.F., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. 2022. Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. Cell **185**(24): 4574-4586.e16. doi:10.1016/j.cell.2022.10.020.

Bhattarai, B., Bhattacharjee, A.S., Coutinho, F.H., and Goel, R.K. 2021. Viruses and Their Interactions With Bacteria and Archaea of Hypersaline Great Salt Lake. Frontiers in Microbiology **12**. Available from https://www.frontiersin.org/articles/10.3389/fmicb.2021.701414 [accessed 14 March 2023].

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5): 1792–1797. doi:10.1093/nar/gkh340.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**(12): 1647–1649. doi:10.1093/bioinformatics/bts199.

Mehta, D., and Satyanarayana, T. 2016. Bacterial and Archaeal α-Amylases: Diversity and Amelioration of the Desirable Characteristics for Industrial Applications. Front Microbiol **7**: 1129. doi:10.3389/fmicb.2016.01129.

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. 2020. CRISPR-CasΦ from huge phages is a hypercompact genome editor. Science **369**(6501): 333–337. doi:10.1126/science.abb1400.

Pinilla-Redondo, R., Russel, J., Mayo-Muñoz, D., Shah, S.A., Garrett, R.A., Nesme, J., Madsen, J.S., Fineran, P.C., and Sørensen, S.J. 2022. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. Nucleic Acids Res. **50**(8): 4315–4328. doi:10.1093/nar/gkab859.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033.