

Supporting Information for Fischer et al., "Giant virus with a remarkable complement of genes infects marine zooplankton" doi:10.1073/pnas.1007615107

1. Supporting Materials and Methods

2. Supporting References

3. Supporting Tables 1-3

4. Supporting Figures 1-15

1. Supporting Materials and Methods

CroV purification

Flagellate growth was monitored by staining cells with Lugol's Acid Iodine and counting by microscopy using a hemocytometer, which had a detection limit of 1×10^3 cells/ml. Lysates were centrifuged for 1 hour at 10,500 x g in a Sorvall RC-5C centrifuge (GSA rotor, 4°C) to remove most bacteria and cell debris. The supernatant was centrifuged for 1 hour at 150,000 x g in a Sorvall RC80 ultracentrifuge (SW40 rotor, 20°C). Pelleted material from ultracentrifugation was not immediately resuspended, but pellets from four to five consecutive ultracentrifuge runs were stacked for increased virus concentration. Pellets were resuspended in 0.2 - 0.5 ml sterile 50 mM Tris-HCl, pH 7.6, loaded onto a 20/30/40/50% (wt/vol in 50 mM Tris-HCl, pH 7.6) sucrose gradient, and centrifuged for 1 hour at 70,000 x g in a Sorvall RC80 ultracentrifuge (SW40 rotor, 20°C). The 29-36% sucrose fraction, containing the bulk of CroV particles, was extracted from the gradient by pipetting, diluted 1:1 with sterile 50 mM Tris-HCl, pH 7.6, and centrifuged for 1 hour at 150,000 x g (SW40 rotor, 20°C). Virus pellets were resuspended in sterile 50 mM Tris-HCl, pH 7.6 and stored at 4°C. Glutaraldehyde-fixed (0.5% wt/vol) virus was quantified by epifluorescence microscopy (SYBR Green I, Invitrogen; Whatman Anodisc filter membranes, VWR Canada).

CroV DNA extraction

Purified CroV particles were suspended in L buffer (0.01 M Tris-HCl, pH 7.6, 0.1 M EDTA, 0.02 M NaCl) containing 1% (wt/vol) N-lauroylsarcosine and 1 mg/mL Proteinase K and incubated at 55°C for 12 hours. DNA was extracted with equal volumes of phenol (once), phenol/chloroform (1:1, once) and chloroform/isoamylalcohol (24:1, twice). The DNA was precipitated with 0.06 volumes of 5 M NaCl and 2 volumes of -20°C cold 100% ethanol. After centrifugation, the DNA pellet was washed with 70% ethanol, air-dried, and dissolved in nuclease-free molecular grade water (Invitrogen, Burlington, ON, Canada).

Genome sequencing and assembly

High-throughput pyrosequencing on a GS 20 platform (454 Life Sciences, Branford, CT, USA) of 5.4 µg CroV DNA resulted in 543,864 individual sequence reads with a run size of 64.5 Mbases and an average read length of 119 bp. For de novo assembly, Newbler™ Assembler software (454 Life Sciences) was used to generate 49 large contigs, 716-65,787 bp in length. The average sequence coverage of the contigs was 39-fold. The 48 large GS20 contigs that were associated with CroV comprised 592,883 bp of non-redundant sequence with an average contig size of 12,352 bp. Additional pyrosequencing was performed on a GS FLX platform with Titanium chemistry (McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada) using 7.0 µg of CroV DNA. Pyrosequencing on 1/8 of a picotiter plate resulted in 74,111 individual sequence reads with a total data volume of 27.4 Mbases and an average read length of 370 bp. GS De Novo Assembler Software created 44 large contigs, 504-112,465 bp in length, with an average 38-fold coverage. The 38 large GS FLX contigs that were associated with CroV comprised 601,547 bp of non-redundant sequence with an average contig size of 15,797 bp.

To span the inter-contig regions, several oligonucleotide primers were designed for each contig end, their 3' ends distally oriented. Different primer combinations were used in multiplex polymerase chain reactions (PCR) and resulting products were sequenced at the University of British Columbia's Nucleic Acid and Protein Service Facility (Vancouver, BC, Canada) using BigDye V3.1 chemistry.

In addition, alternative assemblies were created (Sequencher v4.8, Gene Codes, Ann Arbor, MI, U.S.A.). Any predicted contig connections were tested by PCR and, if a distinct PCR product was obtained, confirmed by sequencing. A list of primer sequences and PCR conditions is available upon request to the authors. Several regions of the final genome assembly, mainly those containing repeats, were re-sequenced to increase coverage.

A small insert shotgun library was created to aid in the sequencing and assembly of the tRNA gene cluster, as these sequences were absent from 454 contigs, but were overrepresented in clone libraries. Thirty microliters of 150 ng/μl CroV genomic DNA were added to 750 μl of 10% (wt/vol) glycerol in TE, pH 8.0, and sheared by nebulization according to the manufacturer's instructions (Invitrogen, Burlington, ON, Canada). The sheared DNA was RNaseI treated (NEB, Canada), end-repaired (DNATerminator Kit, Lucigen, Middleton, WI, USA), separated on an agarose gel and the 1-5 kb size fraction was extracted. Blunt-ended fragments were ligated into the pSMART-LCKan vector (Lucigen). Plasmids from 288 recombinant *Escherichia coli* clones were isolated and bi-directionally sequenced.

The creation of large insert libraries (pCC1FOS vector [≈40 kb insert size], CopyControl Fosmid Library Production Kit, Epicentre, Madison, WI, USA; pJAZZ-KA vector [10-20 kb insert size], Lucigen) was unsuccessful.

Chromosome length was analyzed by pulsed-field gel electrophoresis (PFGE). Approximately 5×10^8 virions (purified by density gradient centrifugation) were embedded in 1% (wt/vol) low-melting agarose (Invitrogen) in L buffer (0.01 M Tris-HCl, pH 7.6, 0.1 M EDTA, 0.02 M NaCl). The gel plug was incubated in L buffer + 1% (wt/vol) lauroylsarcosine + 1 mg/ml Proteinase K, at 50°C overnight to release the viral genome from the capsid. The gel plug was then washed three times with TE, pH 7.6 for 30 minutes each and once with 0.5x TBE and sealed in a 1% (wt/vol) agarose gel in 0.5x TBE. PFGE was performed for 25 hours at 200 V (6 V/cm), 14°C, 60-120 sec switch time at 120° angle and a ramping factor of 1 6.17 using a BioRad DR2 CHEF unit. Analysis of the genome conformation and verification of the final sequence assembly was done by whole-genome restriction digests with *Fspl*, *Apal*, and *SacII* (all from NEB Canada), followed by PFGE separation of fragments. For each restriction analysis, $\approx 5 \times 10^9$ virions were embedded in low-melting agarose and processed as described above. After the first TE washing step, the plugs were washed twice in TE, pH 7.6, 1 mM phenylmethylsulfonyl fluoride (PMSF) for 1 h, once in TE, pH 7.6 for 30 min, and once in the respective NEB restriction enzyme buffer for 30 min on ice. The gel plugs were then added to 0.2 ml of the respective restriction digest buffer, supplemented with 100 μg/ml bovine serum albumin (BSA) and 20 units of the respective restriction enzyme. After 20 min incubation at room temperature, the gel plugs were incubated at the recommended temperature for 6 hours and then for another 8 hours in a fresh reaction mixture. Following digestion, gel plugs were incubated for 2 hours at 50°C in 300 μl of L buffer with 1 mg/ml Proteinase K. Gel

plugs were rinsed three times with 1x TBE and sealed in a 1% agarose gel. PFGE run conditions varied according to DNA fragment lengths.

Genome annotation

Artemis software v12.0 (1) was used for genome annotation. CDSs predicted by Artemis were compared to those predicted by the EMBOSS application GETORF (2). We defined a CDS as being initiated by a start codon and terminated by a stop codon, with a minimum length of 50 uninterrupted consecutive codons (with the exception of crov299a). CDSs overlapping with a larger CDS or exhibiting a strongly biased amino acid composition were removed. Alternative start codons (ATA, ATT) were used to initiate a few apparent CDSs that lacked an initiator Met. This was the case for crov004, crov025, crov066, crov070, crov158, crov184, crov251, crov258, crov348, crov355, crov438, crov441, crov511, crov521, and crov524. Translated CDSs were searched against the NCBI non-redundant (nr) database using BLASTP (3) with a conservative E-value cutoff of 1e-05 to avoid contamination by false homologs, and the COG database (4) was searched using the NCBI BLAST option 'search for conserved domains'. Functional annotation resulted from integrating BLAST results with conserved protein domains identified via the Pfam (5) and InterPro (6) databases. In cases where these predictions were still ambiguous or inconclusive, multiple sequence alignments with putative homologs were created to infer functional predictions (e.g. crov492, Rpb9). For NCVOG analysis, a BLASTP search of CroV CDSs was conducted against a database containing all NCLDV proteins used by Yutin et al. (7) (downloaded from <ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG>). Hits with E-values below 1e-05 were assigned to their respective NCVOGs. Putative tRNA genes were identified with tRNAscan-SE using the general tRNA model (8); codon analysis was carried out using CodonW (<http://mobylye.pasteur.fr/cgi-bin/MobylyePortal/portal.py?form=codonw>).

Before calculation of the average size of intergenic regions, the two-tailed 5% most extreme data points were trimmed off. Promoter analysis was carried out by examining the 100-nt regions immediately upstream of CroV CDSs using MEME (10). MEME analysis returned the putative early promoter motif with the consensus sequence "AAAAATTGA". The position of this motif relative to the start codon was defined as the number of nucleotides between the first adenine in AAAAATTGA and the first nucleotide of the predicted start codon. Next, we searched for a potential late promoter motif in the 100-nt upstream regions of selected CDSs predicted to encode structural components: major capsid protein (crov342), major core protein (crov332), capsid protein 2 (crov398), capsid protein 3 (crov321), capsid protein 4 (crov176), and a phage tail collar domain-containing protein (crov148). With the exception of crov176, all these CDSs were preceded by a perfectly conserved "TCTA" motif that was flanked by AT-rich sequences (containing up to 1 G or C in the 11 nt upstream of TCTA and up to 3 G or C in the 10 nt downstream of TCTA). The TCTA motif was located 11-20 nt upstream of the predicted start codon (as defined by the number of nucleotides between the first thymidine of TCTA and the first nucleotide of the predicted start codon). Based on this sequence profile, we examined the 30-nt upstream region of the 124 late genes for further occurrences of the motif using MEME. Consensus sequence logos were created with WebLogo (11).

Phylogenetic analysis

For phylogenetic reconstruction, putative homologs of the query protein were identified by separate BLAST searches against GenBank nr databases of viruses, eukaryotes, bacteria, and archaea, or, where necessary, taxonomic subgroups thereof. Upon visual inspection of the potential homologs, a representative set of sequences was selected for further analysis. Alternatively, some sequences were downloaded directly via their GenBank accession numbers or keyword searches.

Multiple sequence alignments were created using MUSCLE (12), followed by manual refinement. Bayesian Inference (BI) analysis as implemented in MrBayes v3.1.2 (13) was carried out using the following settings: rates=gamma, aamodelpr=mixed.

MrBayes was run for at least 1 million generations or until the standard deviation of split frequencies was less than 0.01. BI trees were generated by the majority rule consensus method. The phylogeny.fr server was used for Maximum Likelihood analysis (14).

Microarray analysis

To create the microarray, oligonucleotides 50-70 bp in length were designed for 438 of the 544 predicted CroV CDSs. The oligonucleotide probes were printed onto amino silane treated glass slides using a BioRobotics MicroGrid 2 printer. Each virus-specific probe was printed in five replicates along with several negative and positive control probes.

RNA extraction

Total RNA was extracted from host cells that had been infected with CroV at an MOI of ≈ 2 as well as from an uninfected control culture. The uninfected control hybridization consisted of one biological replicate and five technical replicates, and the sole purpose of hybridizing mRNA from uninfected cultures to the CroV microarray was to detect cases where host mRNA cross-hybridized with the virus-specific probes.

Six flasks each containing 600 ml of an exponentially growing *C. roenbergensis* culture were infected with CroV lysate at a cell density of 1×10^5 per ml. Subsamples of 300 ml were taken at T=0, 1, 2, 3, 6, 12, 24, 48, and 72 h p.i.. It should be noted that although the CroV infection cycle lasts 12-18 hours, cultures frequently contained living cells up to 5 days post infection. This is due to the low MOI used, which will require more than one round of infection to lyse all cells in the culture. *C. roenbergensis* cells were pelleted by centrifugation (1,500 x g, 15 min, 20°C, Eppendorf A-4-62 rotor), pellets were washed in 2 x 40 ml PBS and centrifuged again. Cells were then resuspended in 2 ml RNeasy lysis solution (Qiagen, Mississauga, ON, Canada) and stored at -80°C until further use. RNA extraction was performed using an RNeasy Protect Midi Kit (Qiagen). Each sample was split into two RNase-free 2 ml microfuge tubes, centrifuged (12,000 x g, 5 min) and the pellets resuspended in 2 x 2 ml RLT buffer containing 20 μ l β -mercaptoethanol. After vortexing 10 times for 10 sec each, samples were centrifuged (20,000 x g, 5 min) and the supernatant was transferred to a 15 ml Falcon tube containing 4 ml of 70% ethanol. Following vigorous shaking the samples were applied to an RNeasy Midi column, centrifuged (3,220 x g, 10 min, 22°C) and the flow-through was discarded. This process was repeated once until the entire sample had been applied to the column. Columns were washed once with 4 ml RW1

buffer (3,220 x g, 5 min), twice with RPE buffer (3,220 x g, 5 min) and transferred to a new Falcon tube. To elute the RNA, 250 µl RNase-free water was added to the column; samples were incubated at room temperature for 1 min and centrifuged (3,220 x g, 5 min). The elution process was repeated once and both eluates were combined. RNA was precipitated by adding 250 µl of 7.5 M NH₄Ac and 1 ml of 100% ethanol and incubating the samples at -80°C overnight. Following centrifugation (20,000 x g, 30 min), the pellet was washed twice with 0.5 ml 80% ethanol (20,000 x g, 30 min). The pellet was air dried, resuspended in 50 µl RNase-free water and stored at -80°C. RNA quantity and quality was assessed using the Agilent Bioanalyzer 2100 system (www.agilent.com).

DNase treatment of total RNA

10 µl of total RNA, 2.5 µl of Turbo DNase buffer (10x, Ambion, UK) and 2.5 µl of Turbo DNase (2 U/µl, Ambion, UK) were combined in a total volume of 25 µl and incubated at 37°C for 15 min. Following the addition of 5 µl DNase inactivation reagent 8174G (Ambion, UK) and mixing, the samples were incubated at room temperature for 3 min, centrifuged (14,000 x g, 1 min), and the supernatant was transferred to a new RNase-free microfuge tube.

cDNA synthesis

The Microarray Target Amplification Kit (Roche, UK) was used for cDNA synthesis. For each of the 10 samples, 500 ng total RNA (DNase treated), 0.5 ng spike mRNA (mRNA spikes 1+2, Stratagene, UK), and 1 µg TAS-T7 Oligo dT were combined in a total volume of 10.5 µl, mixed briefly, and incubated at 70°C for 10 min. A reaction mix containing 4 µl 5x first strand buffer, 2 µl 0.1 M DTT, 2 µl 10 mM dNTP mix, and 1.5 µl reverse transcriptase (17 U/µl) was added and samples were incubated at 42°C for 2 hours followed by 95°C for 5 min and cooling on ice. For second strand synthesis, a reaction mix was added to a final volume of 50 µl containing 2.5 µl dNTP mix (10 mM), 5 µl TAS-(dN)₁₀ primer (100 µM), 5 µl Klenow Reaction Buffer (10x), and 4 µl Klenow enzyme (2 U/µl). After brief mixing, the reaction was incubated at 37°C for 30 min. Following the addition of 1.25 µl carrier RNA (0.8 µg/µl) and 50 µl RNase-free water, cDNA was purified using the Microarray Target Purification Kit (Roche, UK) according to the manufacturer's instructions. cDNA was PCR-amplified using the following reaction setup: 12.5 µl purified ds cDNA, 1 µl TAS primer (50 µM), 2 µl dNTP mix (10 mM), 10 µl Expand PCR buffer (10x), 1.5 µl Expand enzyme mix (3.5 U/µl), 73 µl RNase-free water. PCR conditions were as follows: one cycle of 2 min at 95°C and 24 cycles of 30 sec at 95°C, 30 sec at 55°C, 3 min at 72°C. PCR products were purified using the Microarray Target Purification Kit (Roche) according to the manufacturer's instructions and concentrated on a Microcon YM-30 column (Millipore, UK) to a final volume of 10.75 µl.

Labeling of cDNA with fluorescent dyes

PCR-amplified cDNA was labeled with Cy3 by *in vitro* transcription using the Microarray Target Synthesis Kit (Roche, UK). 10.75 µl of template DNA were combined with 2 µl DTT (100 mM), 1 µl NTP mix (25 mM ATP, 25 mM CTP, 25 mM GTP, 18.75 mM UTP), 1.25 µl Cy3-17-UTP (5 mM, Amersham, UK), 2 µl transcription

buffer (10x) and 3 μ l transcription enzyme blend. The reaction was incubated for 16 hours at 37°C and Cy3-labeled cRNA was purified using the Microarray Target Purification Kit (Roche, UK) according to the manufacturer's instructions.

Microarray hybridization and data analysis

Prior to hybridization, glass slides were incubated at 42°C for 3 hours with gentle agitation in a solution containing 1% BSA (PAA Laboratories, UK), 5x SSC (Sigma, UK), and 0.1% SDS. For hybridization, 9 μ l 20x SSC, 1.2 μ l 10% SDS, and the labeled cRNA samples were combined in a total volume of 60 μ l and prewarmed to 60°C. Samples were loaded onto the microarray slide covered by a Lifterslip (Erie Scientific Company, UK) and hybridization was performed in a microarray hybrid chamber (Camlab, UK) at 60°C for 20 hours. Microarray slides were scanned using an Affymetrix 418 Array Scanner with GMS Scanner software v1.51.0.42. Scans were performed at 10 gain increments to determine the optimal scanning range for signal distribution. CroV genomic DNA labeled with Cy3-dCTP (GE Healthcare, UK) was used to test the microarray. To assign a preliminary transcription activity status to each CroV gene, probe spots were individually assessed using a manual scoring system (15) performed on the original microarray images (ImaGene 5.6.1, BioDiscovery, UK). In order to separate signal from background noise, normalized fluorescence signals were plotted with increasing values to yield an intensity distribution plot such as the one shown in Fig. S15. The signal threshold was then set manually in a region where the intensity values started to increase exponentially. A CDS was considered to be expressed if an above background signal was detected in at least 3 of the 5 replicate spots within an array of one of the 9 time points and if the respective spot did not produce a signal when hybridized with labeled cRNA isolated from the uninfected control culture. Very few genes belonged to the 3/5 category and only four of them (crov045, crov062, crov220, crov223) were considered to be expressed based on a 3/5 condition. Microarray experimentation and data was collected to be MIAME compliant.

Host strain 18S Sequencing

Eukaryotic 18S rDNA fragments were amplified from *C. roenbergensis* using universal eukaryotic primers Euk1A and Euk516r as previously described (16). PCR products were cloned into the pCR4-TOPO vector (Invitrogen, Burlington, ON, Canada) and sequenced at the University of British Columbia's Nucleic Acid and Protein Service Facility (Vancouver, BC, Canada) using BigDye V3.1 chemistry.

2. Supporting References

1. Rutherford K et al. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16:944-945.
2. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276-277.
3. Altschul SF et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
4. Tatusov RL et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29:22-8.
5. Bateman A et al. (2002) The Pfam protein families database. *Nucleic Acids Res* 30:276-280.
6. Hunter S et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37:D211-5.
7. Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleocytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology* 6:223.
8. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955-964.
9. Shackleton LA, Parrish CR, Holmes EC (2006) Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses. *J Mol Evol* 62:551-563.
10. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.
11. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188-1190.
12. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-7.
13. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.

14. Dereeper A et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36:W465-9.
15. Allen MJ, Martinez-Martinez J, Schroeder DC, Somerfield PJ, Wilson WH (2007) Use of microarrays to assess viral diversity: from genotype to phenotype. *Environ Microbiol* 9:971-982.
16. Diez B, Pedros-Alio C, Marsh TL, Massana R (2001) Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Appl Environ Microbiol* 67:2942-2951.

3. Supporting Tables

Table S1. Novel viral features in the CroV genome.

Predicted Function	CroV CDS	Category
CPD class I photolyase	crov115	DNA replication and repair
Exodeoxyribonuclease VII large subunit	crov048	DNA replication and repair
Exonuclease III / AP endonuclease family 1	crov106	DNA replication and repair
DNA topoisomerase IB, human subfamily	crov152	DNA replication and repair / Transcription
Elp3-like histone acetyltransferase	crov391	Transcription
Eukaryotic translation initiation factor 2 α , SUI2 homolog	crov162	Translation
Eukaryotic translation initiation factor 2 γ	crov479	Translation
Eukaryotic translation initiation factor 5B	crov113	Translation
Isoleucyl-tRNA synthetase	crov505	Translation
tRNA pseudouridine 5S synthase	crov071	Translation
Bifunctional 3-deoxy-D- <i>manno</i> -2-octulosonate 8-P phosphatase / arabinose 5-phosphate isomerase	crov265	Lipopolysaccharide biosynthesis
Bifunctional N-acylneuraminatase cytidyltransferase / demethylmenaquinone methyltransferase	crov266	Lipopolysaccharide biosynthesis
Bifunctional 3-deoxy-D- <i>manno</i> -2-octulosonate 8-P synthase / dTDP-6-deoxy-L-hexose 3-O- methyltransferase	crov267	Lipopolysaccharide biosynthesis
Cysteine dioxygenase type I	crov413	Sulfate production
Ubiquitin-activating enzyme E1	crov435	Protein modification
Intein insertions in DNA polymerase B, DNA topoisomerase IIA, Ribonucleoside-diphosphate reductase large subunit, RNA polymerase II subunit 2	crov497, crov325, crov454, crov224	Inteins

Table S2. tRNA genes in the CroV genome.

tRNA #	Begin	End	Type	Anti Codon	Cove Score
1	509015	509086	Tyr	GTA	67.17
2	509181	509262	Leu	TAA	64.77
3	509266	509333	Ser	CGA	34.68
4	509421	509502	Leu	TAA	63.27
5	509506	509580	Lys	TTT	76.45
6	509587	509658	Sup ochre	TTA	57.71
7	509911	509992	Leu	TAA	64.77
8	509995	510062	Unknown	???	19.09
9	510066	510135	Unknown	???	50.79
10	510177	510258	Leu	TAA	64.77
11	510262	510329	Ser	CGA	29.90
12	510417	510498	Leu	TAA	63.27
13	510502	510569	Ser	CGA	32.05
14	510656	510737	Leu	TAA	63.27
15	510741	510815	Lys	TTT	77.56
16	511091	511172	Leu	TAA	64.77
17	511176	511243	Ser	CGA	29.90
18	511330	511411	Leu	TAA	63.27
19	511415	511482	Ser	CGA	32.05
20	511570	511651	Leu	TAA	63.27
21	511655	511729	Lys	TTT	77.56
22	511736	511809	Asn	GTT	71.75

Table S3. Top BLASTP results for the 34 CDSs in the 38-kb genomic fragment. Predicted bifunctional proteins were split into their N-terminal (NT) and C-terminal (CT) domains and subject to separate BLAST searches.

CroV CDS	Top BLASTP hit	Accession number	E-value	Amino acid identity	Alignment length (aa)
crov242	UDP-glucose 6-dehydrogenase [<i>Bacteroides</i> sp. D4]	ZP_04557566	9e-23	32%	243
crov243	DNA integration/recombination/inversion protein [<i>Helicobacter bilis</i> ATCC 43879]	ZP_04581560	0.83	41%	65
crov244	hypothetical protein RB2150_01259 [<i>Rhodobacterales bacterium</i> HTCC2150]	ZP_01742127	7e-06	28%	164
crov245	-	-	-	-	-
crov246	alpha-2,3-sialyltransferase [<i>Campylobacter coli</i> RM2228]	ZP_00368088	6e-06	32%	143
crov247	glycosyltransferase family 2 [<i>Cyanothece</i> sp. PCC 7424]	YP_002381075	3e-07	37%	108
crov248	hypothetical protein Phep_1979 [<i>Pedobacter heparinus</i> DSM 2366]	YP_003092249	4.1	36%	52
crov249	-	-	-	-	-
crov250	hypothetical protein Swol_1940 [<i>Syntrophomonas wolfei</i> subsp. <i>wolfei</i> str. Goettingen]	YP_754608	1e-11	26%	298
crov251	chromosomal replication initiator protein DnaA [<i>Leptospira biflexa</i> serovar Patoc strain 'Patoc 1 (Paris)']	YP_001837424	3.7	31%	73

CroV CDS	Top BLASTP hit	Accession number	E-value	Amino acid identity	Alignment length (aa)
crov252	unknown protein [<i>Sphingomonas sp.</i> S88]	AAC44076	3e-09	30%	117
crov253	hypothetical protein FTN_1254 [<i>Francisella tularensis</i> subsp. <i>novicida</i> U112]	YP_898889	0.02	27%	108
crov254	hypothetical protein BACCELL_01591 [<i>Bacteroides cellulosilyticus</i> DSM 14838]	ZP_03677254	7e-12	26%	235
crov255	hypothetical protein [<i>Strongylocentrotus purpuratus</i>]	XP_794970	2.9	34%	100
crov256	conserved hypothetical protein [<i>Bacteroides fingoldii</i> DSM 17565]	ZP_05416596	4e-12	30%	201
crov257	-	-	-	-	-
crov258	-	-	-	-	-
crov259	hypothetical protein NAEGRDRAFT_78502 [<i>Naegleria gruberi</i>]	XP_002681081	5e-14	30%	249
crov260	tetratricopeptide TPR_2 [<i>Arthrospira platensis</i> str. Paraca]	ZP_06380292	6e-35	34%	287
crov261	hypothetical protein GSU3022 [<i>Geobacter sulfurreducens</i> PCA]	NP_954064	8e-41	40%	217
crov262	predicted protein [<i>Thalassiosira pseudonana</i> CCMP1335]	XP_002288935	1e-03	23%	198

CroV CDS	Top BLASTP hit	Accession number	E-value	Amino acid identity	Alignment length (aa)
crov263	hypothetical protein NY2A_B094R [<i>Paramecium bursaria</i> Chlorella virus NY2A]	YP_001497290	3e-20	31%	239
crov264	pyrrolo-quinoline quinone [<i>Conexibacter woesei</i> DSM 14684]	YP_003395319	5.3	33%	57
crov265	NT: CMP-N-acetylneuraminic acid synthetase [<i>Pelobacter carbinolicus</i> DSM 2380]	YP_356697	1e-29	41%	158
	CT: predicted protein [<i>Populus trichocarpa</i>]	XP_002306858	5e-35	31%	297
crov266	NT+CT: cytidyltransferase domain protein [gamma proteobacterium HTCC5015]	ZP_05062599	9e-88	41%	415
crov267	NT: 2-dehydro-3-deoxyphosphooctonate aldolase, putative [<i>Ricinus communis</i>]	XP_002522197	5e-79	57%	254
	CT: hypothetical protein Tery_3108 [<i>Trichodesmium erythraeum</i> IMS101]	YP_722714	1e-29	36%	207
crov268	transposase [<i>Rickettsia</i> endosymbiont of <i>Ixodes scapularis</i>]	ZP_04699603	4.3	27%	102
crov269	hypothetical protein Syncc9605_1741 [<i>Synechococcus</i> sp. CC9605]	YP_382043	5e-21	32%	264
crov270	-	-	-	-	-

CroV CDS	Top BLASTP hit	Accession number	E-value	Amino acid identity	Alignment length (aa)
crov271	tetratricopeptide TPR_2 repeat protein [<i>Arthrospira platensis</i> str. Paraca]	ZP_06381863	6e-13	26%	316
crov272	capsular protein [<i>Haloquadratum walsbyi</i> DSM 16790]	YP_659198	8e-10	30%	213
crov273	glycosyltransferase [<i>Prochlorococcus marinus</i> str. MIT 9215]	YP_001484629	0.23	27%	183
crov274	-	-	-	-	-
crov275	cysteine-rich protein H [<i>Helicobacter pylori</i> HPAG1]	YP_627080	1.5	33%	75

4. Supporting Figures

Amino acid frequency (%)

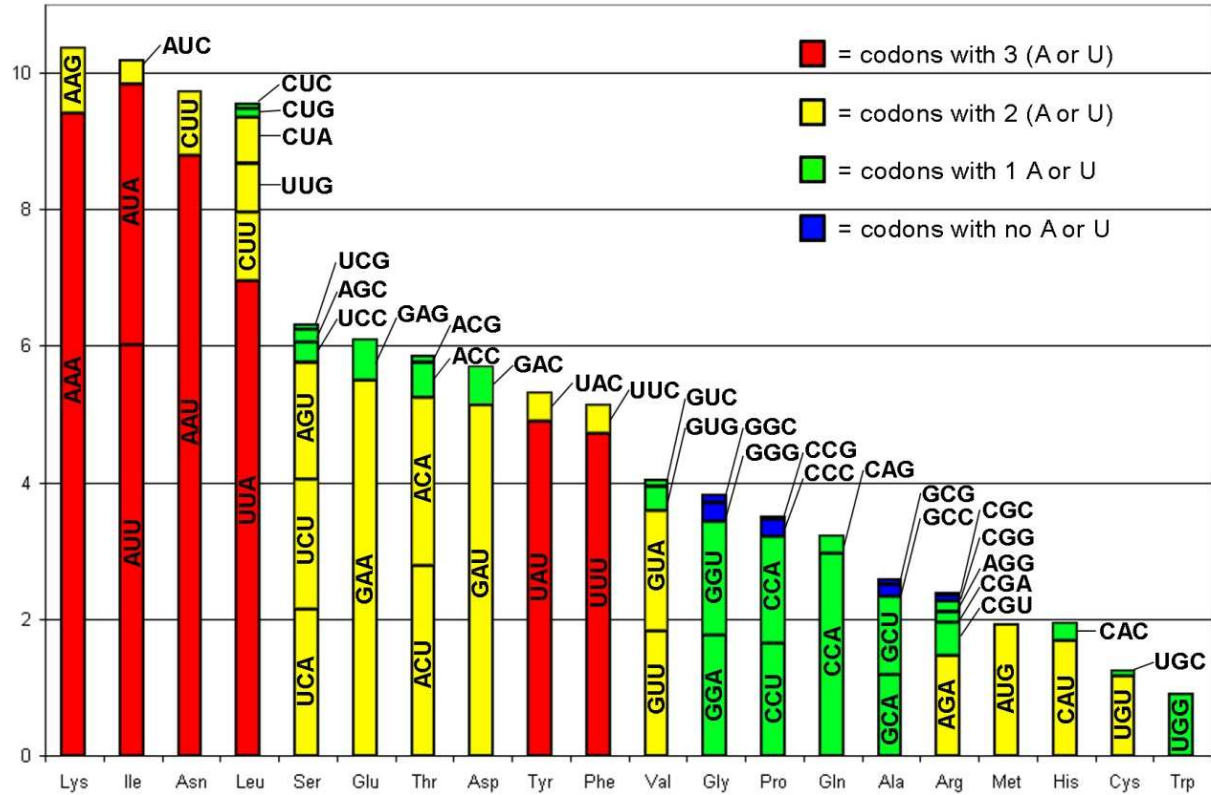


Fig. S1. Codon usage in CroV. This codon analysis is based on 185,006 codons in 544 CDSs (stop codons excluded). Codons consisting 100% of (A or U) are colored in red, 67% (A or U) in yellow, 33% (A or U) in green, 0% (A or U) in blue. The height of each codon column represents the overall frequency of that codon.

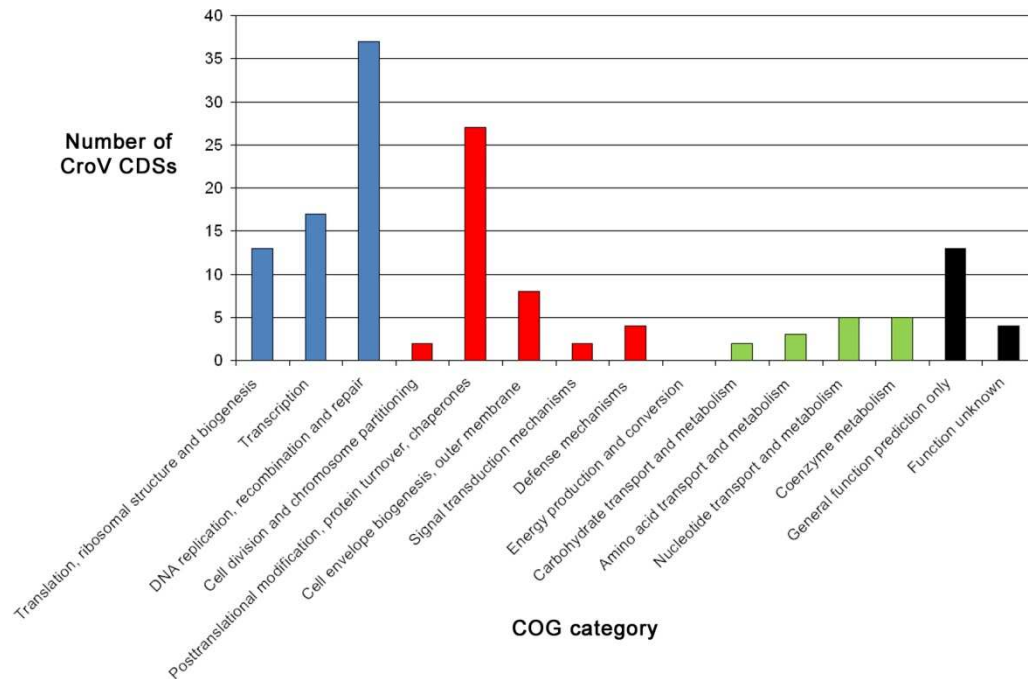


Fig. S2. Functional categories of Clusters of Orthologous Groups of proteins (COGs) identified in CroV. Color code: blue, information storage and processing; red, cellular processes; green, metabolism-related categories; black, poorly characterized categories.

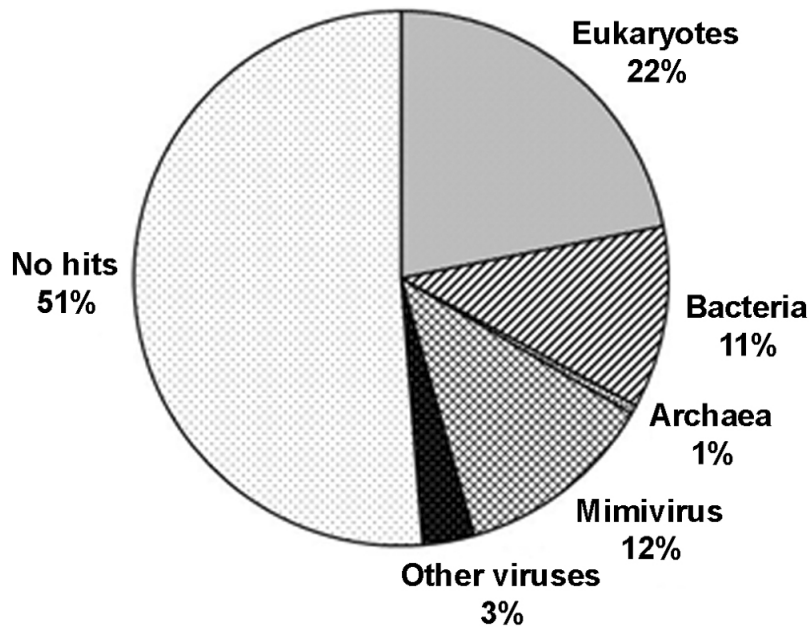


Fig. S3. Distribution of top BLASTP hits for CroV CDSs.

All 544 CroV CDSs were queried against the NCBI non-redundant database and categorized according to the domain affiliation of their top BLASTP hit. The E-value cutoff for this analysis was 1e-05.

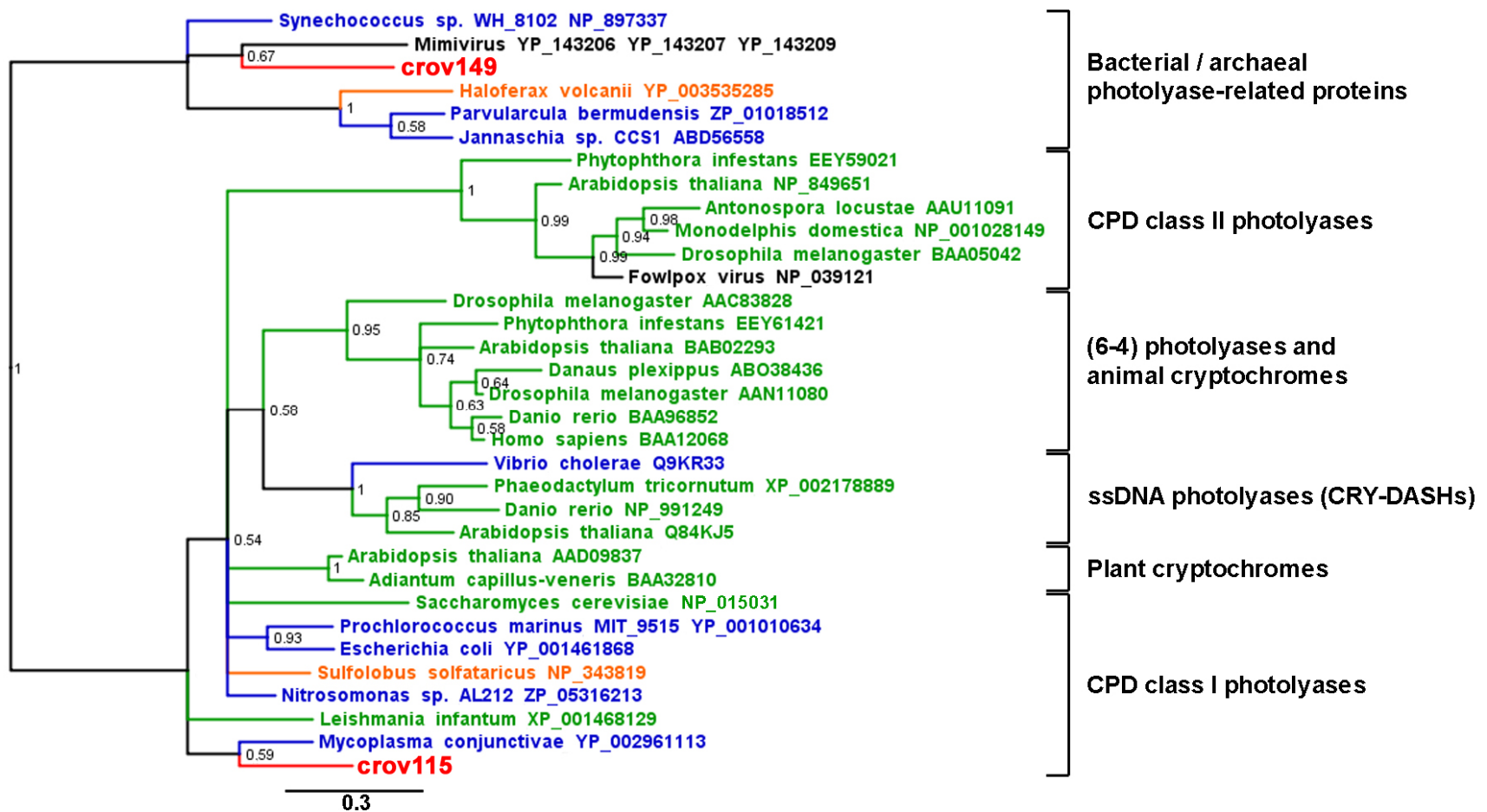


Fig. S4. Phylogenetic analysis of the photolyase/chryptochrome family. The unrooted Bayesian Inference tree of photolyases and chryptochromes is based on 73 conserved sites. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with posterior probabilities and GenBank accession numbers are given for each sequence. The group of bacterial/archaeal photolyase-like proteins belongs to COG3046 (uncharacterized protein related to deoxyribodipyrimidine photolyase), whereas all other groups in this tree belong to COG0415 (deoxyribodipyrimidine photolyase). Due to the low overall sequence conservation among the different groups, the CPD class I photolyases did not resolve into a monophyletic group in this reconstruction.

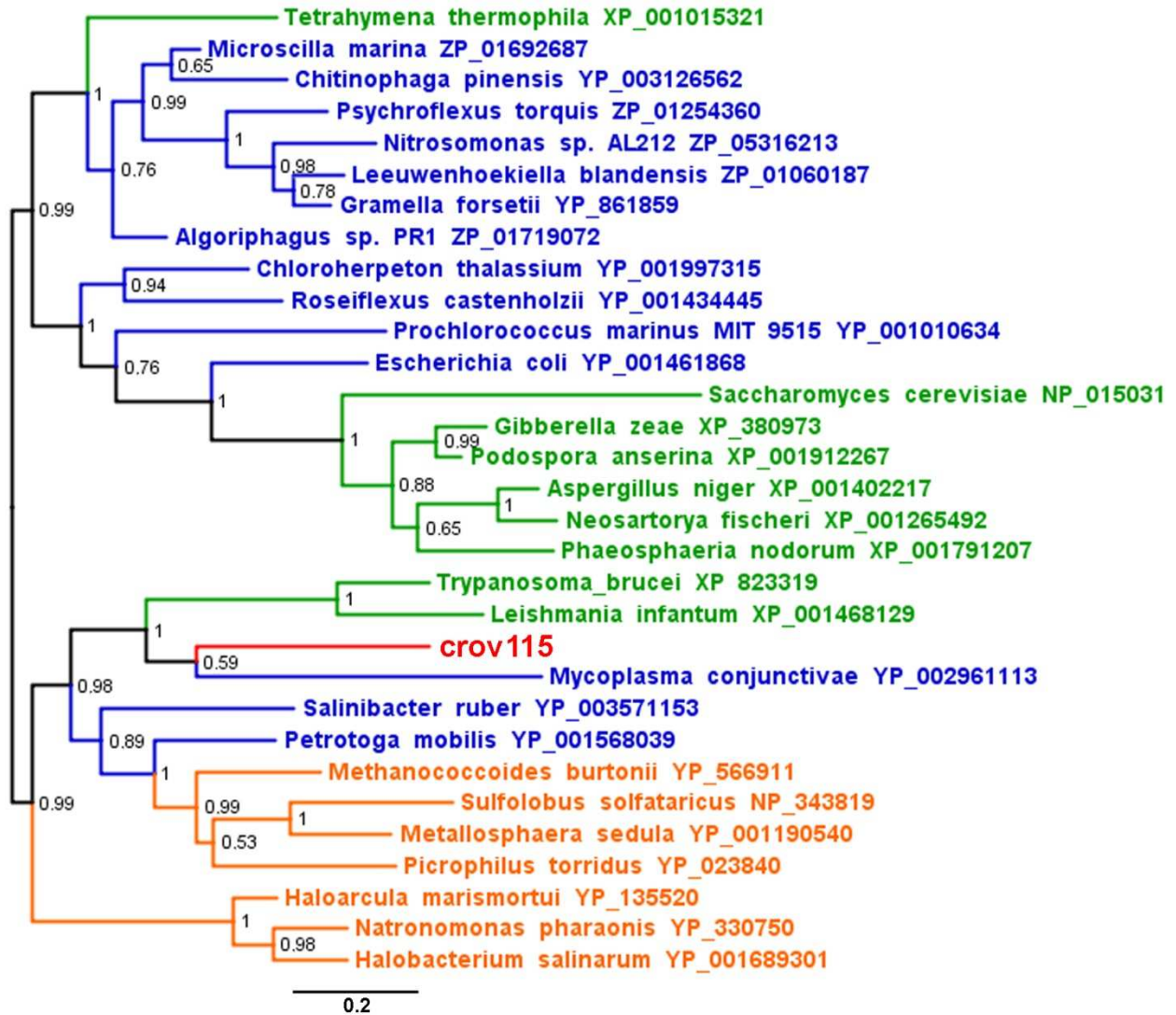


Fig. S5. Phylogenetic analysis of CPD class I photolyases. The unrooted Bayesian Inference tree is based on 189 conserved sites of CPD class I photolyases related to *crov115*. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with posterior probabilities and GenBank accession numbers are given for each sequence.

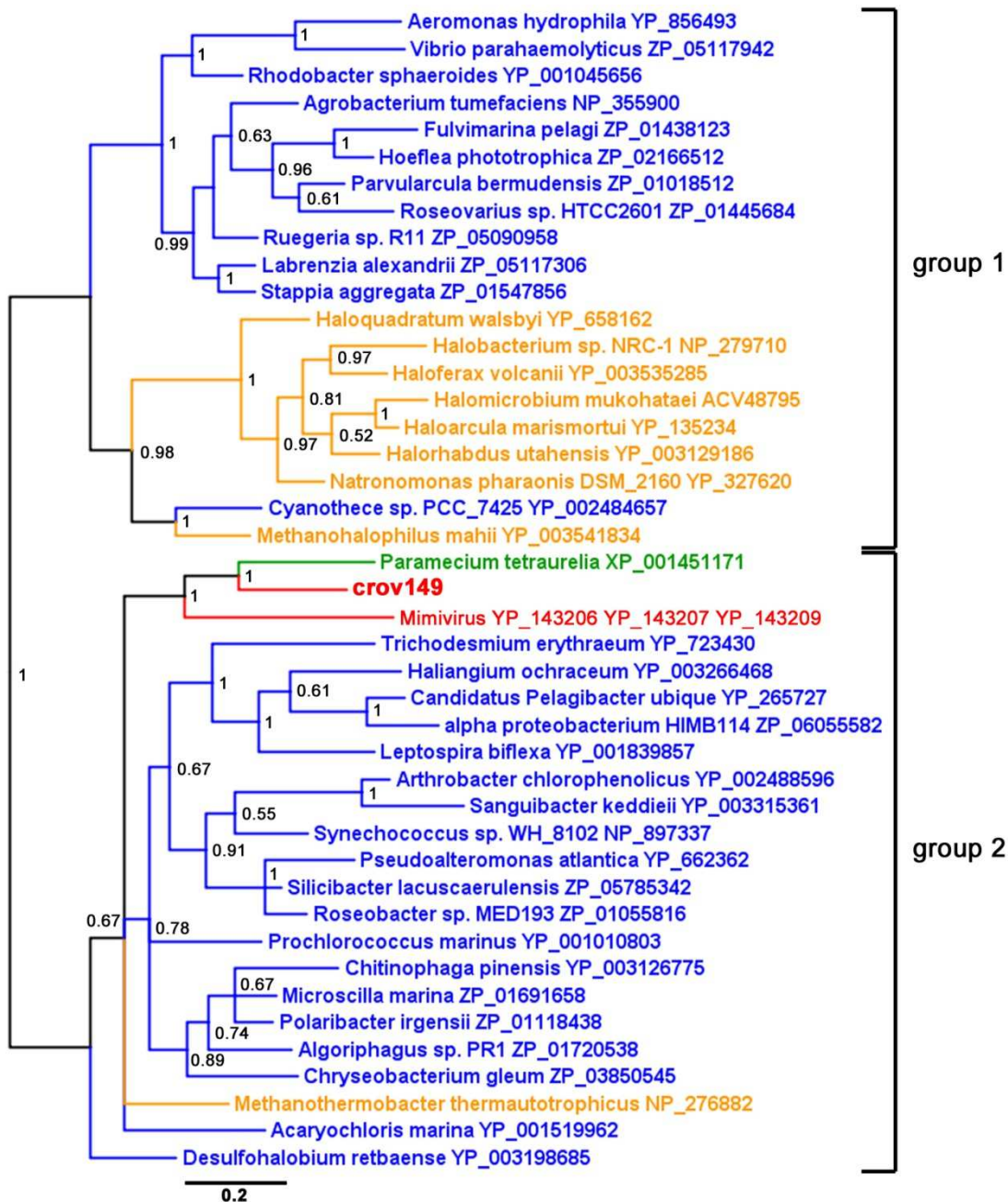


Fig. S6. Phylogenetic position of the predicted DNA photolyase *crov149*. The unrooted Bayesian Inference tree is based on a 157-aa alignment of DNA photolyases related to *crov149*. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Two main groups can be differentiated, group 1 comprising a bacterial and an archaeal clade, and group 2 comprising bacterial and viral sequences. The eukaryotic sequence in group 2 is probably the result of horizontal gene transfer. The Mimivirus photolyase is encoded by three separate CDSs (R852, R853, R855). Nodes are labeled with posterior probabilities and GenBank accession numbers are given for each sequence.

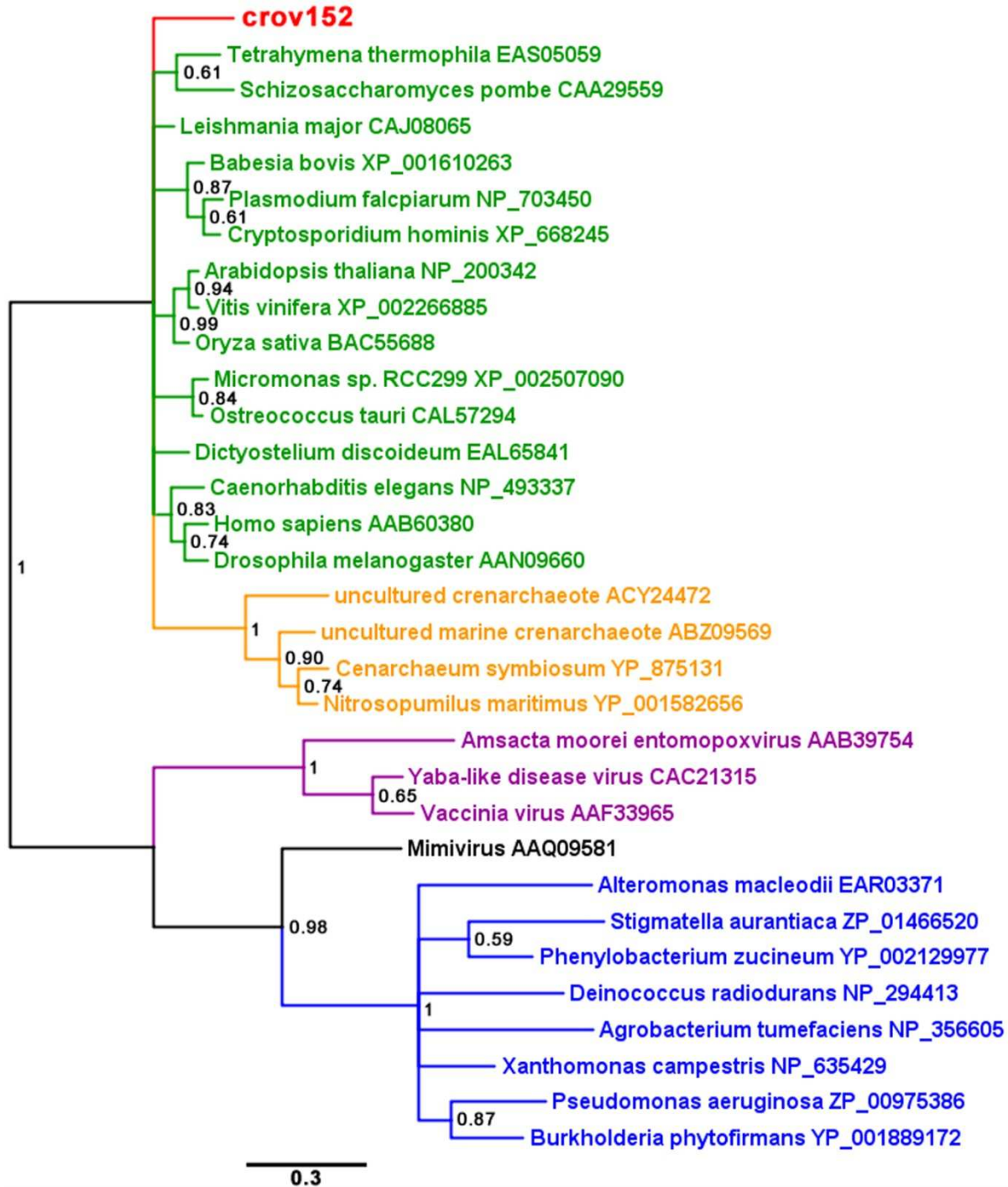


Fig. S7. Phylogenetic position of CroV DNA topoisomerase IB. The unrooted Bayesian Inference tree is based on a 66-aa alignment of DNA topoisomerases of type IB. Sequences are colored orange for archaea, blue for bacteria, green for eukaryotes and purple for poxviruses. Nodes are labeled with posterior probabilities and GenBank accession numbers are given for each sequence.

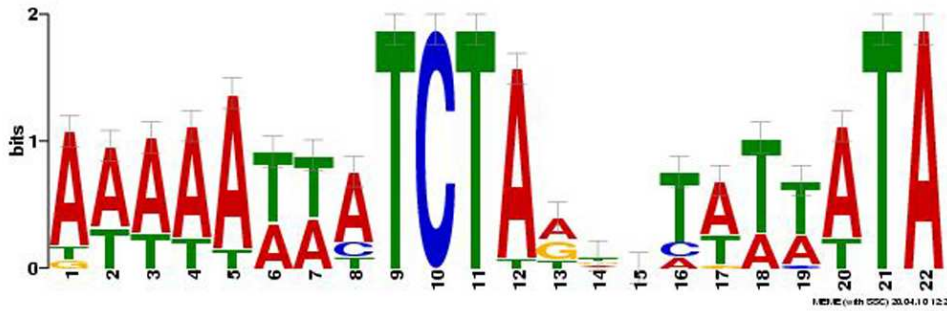


Fig. S8. MEME sequence logo of the CroV late promoter motif.

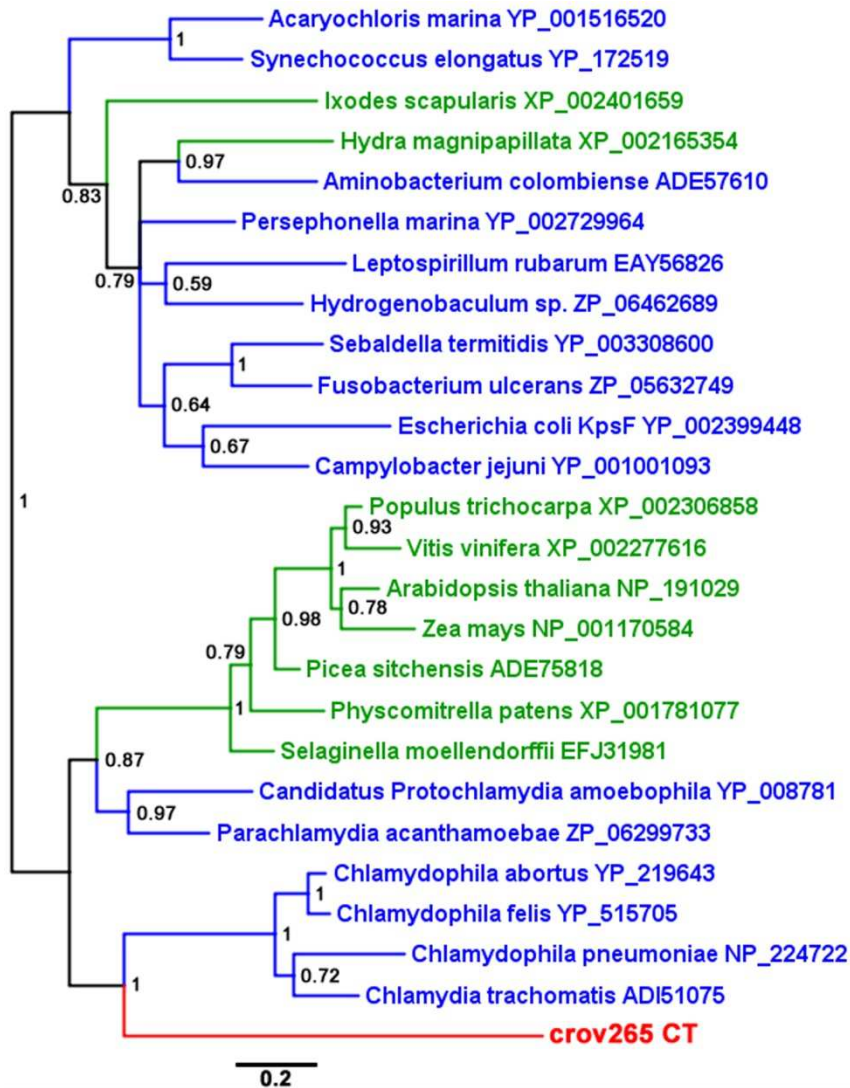


Fig. S9. Phylogenetic tree of API. The unrooted Bayesian Inference tree is based on a 228-aa alignment of arabinose-5-phosphate isomerases (API). Sequences are colored blue for bacteria and green for eukaryotes. Nodes are labeled with support values and GenBank accession numbers are given for each sequence.

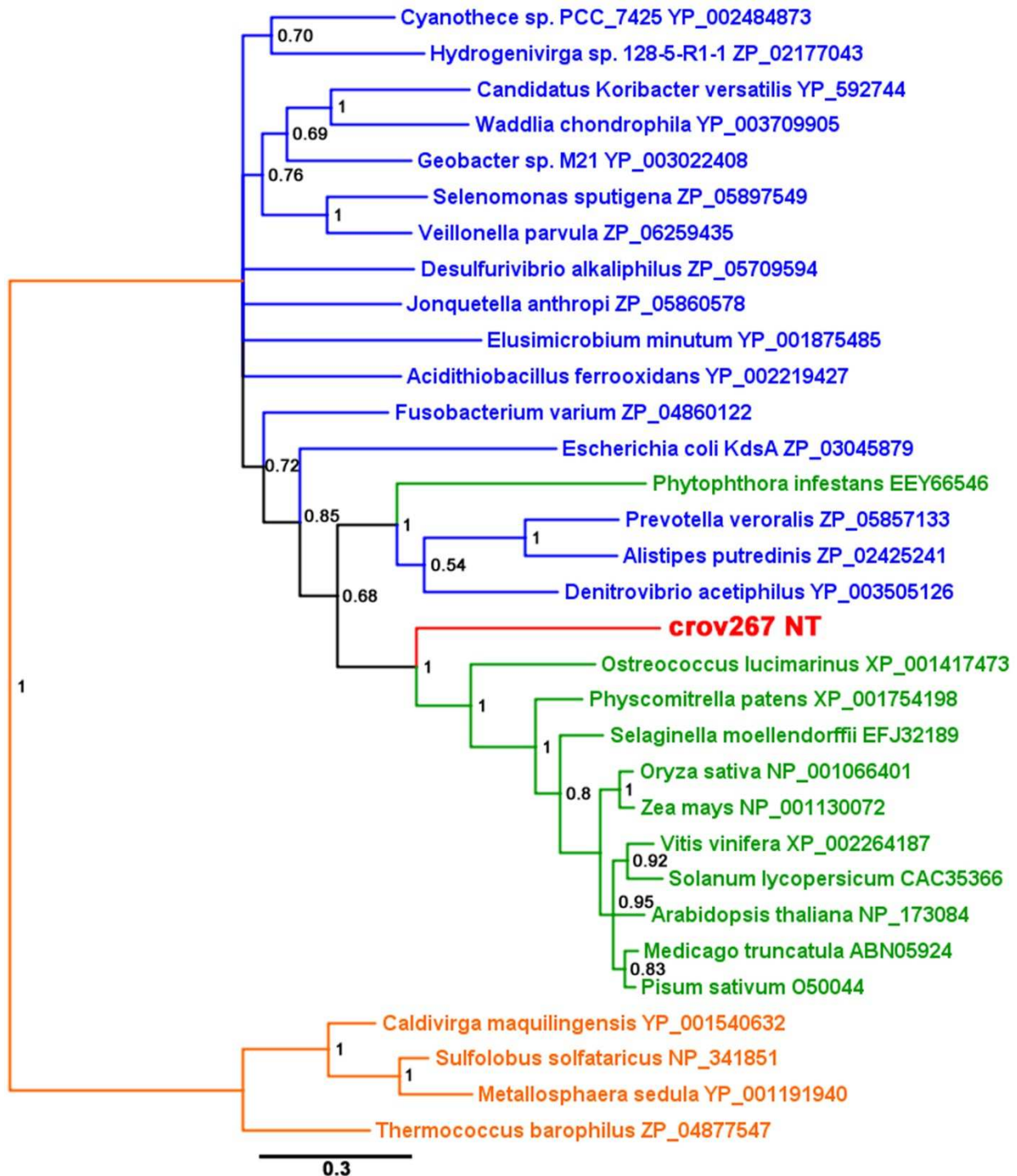


Fig. S10. Phylogenetic tree of KDO 8-P synthases. The unrooted Bayesian Inference tree is based on a 208-aa alignment of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthases (KDOPS). Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with support values and GenBank accession numbers are given for each sequence.

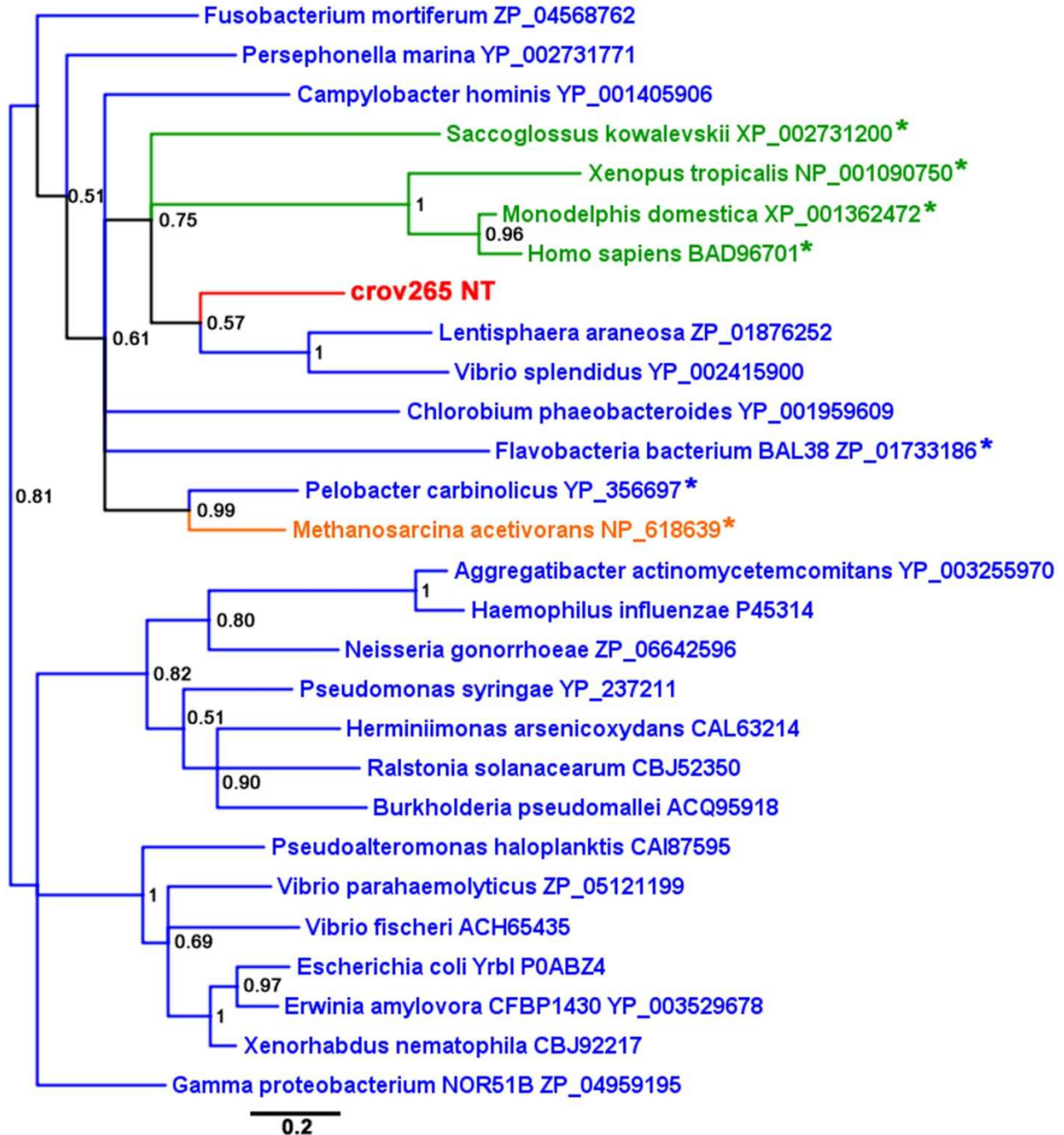


Fig. S11. Phylogenetic analysis of KDO 8-P phosphatases. The unrooted Bayesian Inference tree is based on a 134-aa alignment of 3-deoxy-D-*manno*-octulosonate 8-phosphate phosphatases (KDOPase), which belong to the haloacid dehalogenase-like (HAD) superfamily. Sequences marked with an asterisk are bifunctional enzymes where the C-terminal HAD domain is preceded by a N-acetylneuraminase cytidyltransferase domain. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with support values and GenBank accession numbers are given for each sequence.

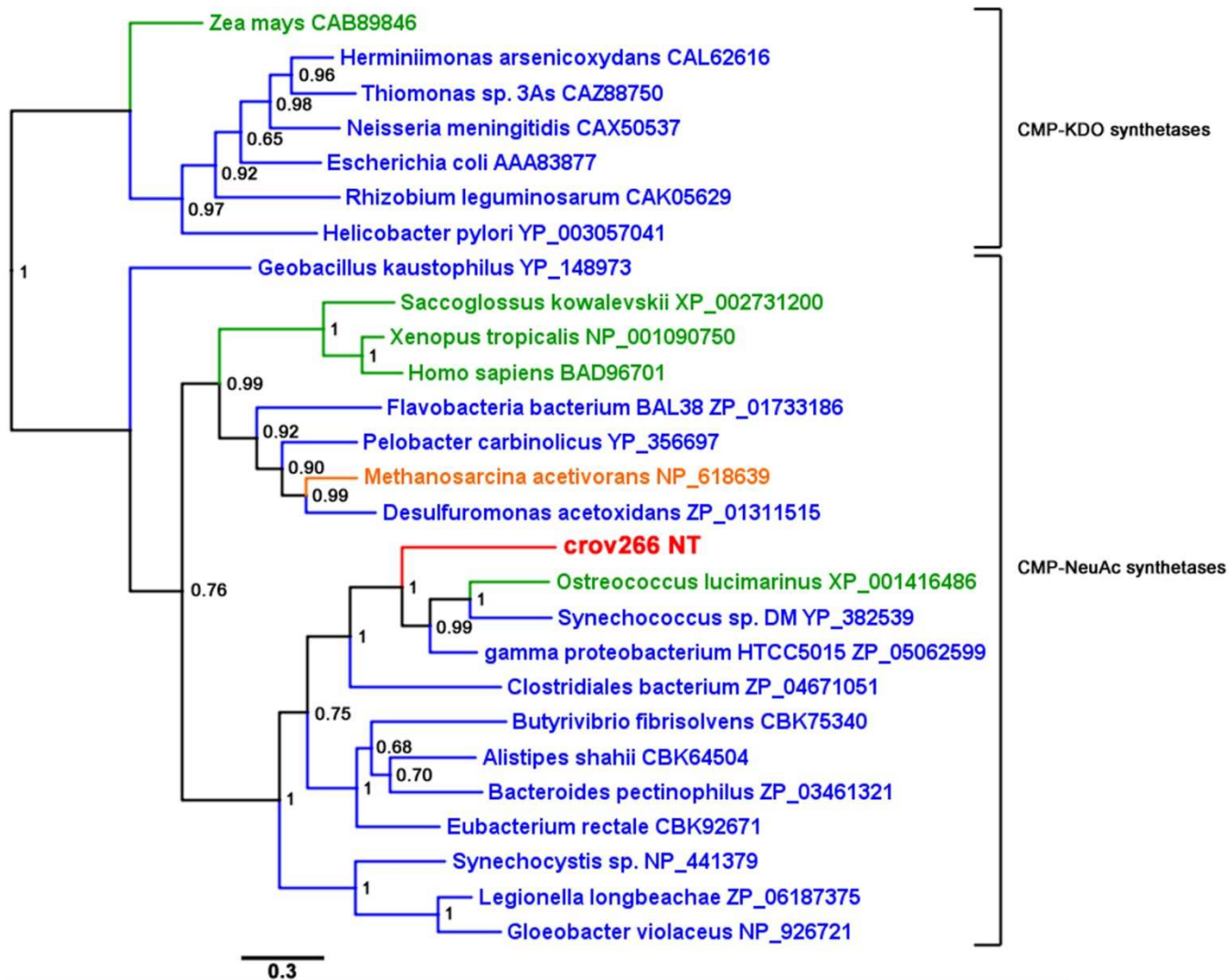


Fig. S12. Phylogenetic tree of two types of cytidyltransferases. The unrooted Bayesian Inference tree is based on a 185-aa alignment of N-acetylneuraminyl cytidyltransferases (CMP-NeuAc synthetases) and 3-deoxy- D-manno - octulosonate cytidyltransferases (CMP-KDO synthetases). Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with support values and GenBank accession numbers are given for each sequence.

Family	Virus	Group I										Group II										Group III														
		VV D5-Type A TPase	DNA Polymerase B	VV A32-Type A TPase	VV A18-Type Helicase	Capsid Protein	Thiol Oxidoreductase	VV D6R-Type Helicase	S/T Protein Kinase	VL TF2-like Transcription Factor	TFII-Like Transcription Factor	MUT-Like NTP Pyrophosphohydrolase	Myristoylated Viron Protein A	Proliferating Cell Nuclear Antigen	Ribonucleotide Reductase, Large Subunit	Ribonucleotide Reductase, Small Subunit	Thymidylate Kinase	dUTPase	RNA Polymerase, Subunit 1	RNA Polymerase, Subunit 2	VL TF3-Like Transcription Factor	RuvC-Like Holliday Junction Resolvase	BroA-Like	Capping Enzyme	ATP-Dependent DNA Ligase	Thioredoxin/Glutaredoxin	SY Phosphatase	BIR Domain	Viron-Associated Membrane Protein	Topoisomerase II	SWI/SNF1 Family Helicase	RNA Polymerase, Subunit 10				
	CroV	crov494	crov497	crov338	crov316	crov342	crov143	crov283	crov309	crov164	crov299	-	-	crov219	crov454	crov452	-	crov069	crov368	crov224	crov341	crov163	-	crov212	-	crov379	-	-	-	crov325	crov402	crov201				
Mimiviridae	Mimivirus	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				
Phycodnaviridae	EhV-86	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				
	EsV-1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
	PBCV-1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
	OtV-1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
	FsV-158	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
Asfarviridae	ASFV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
Iridoviridae	LCDV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
	IIV-6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
unclassified	Marseillevirus	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
Poxviridae	VV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
	MOCV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	AMEV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	MSEV	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Fig. S13. NCLDV core genes found in CroV. Shown are NCLDV core genes of groups I-III present in CroV and selected members of the NCLDV clade. Viral hallmark genes are bolded. *Mimivirus L451 is a putative RuvC-like Holliday Junction Resolvase (HJR) homolog. **OtV-1_053 was identified as a putative RuvC-like HJR homolog

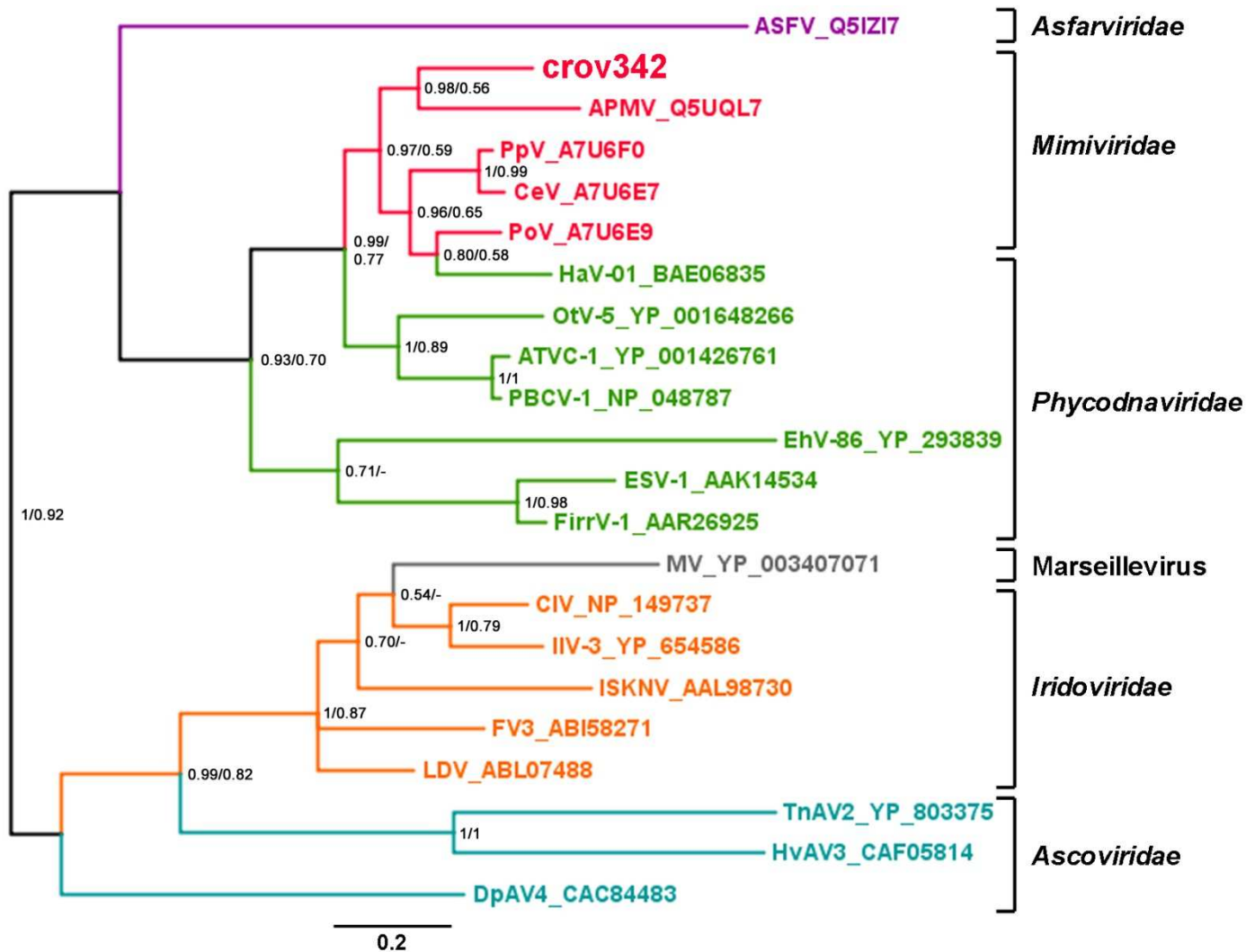


Fig. S14. Phylogenetic analysis of the NCLDV major capsid protein. The unrooted Bayesian Inference tree is based on a 169-aa alignment. Color coding and abbreviations are the same as in Fig. 4. Not included in the tree are the poxviruses, as their capsid proteins are too divergent from those of other NCLDV families. Nodes are labeled with BI posterior probabilities and Maximum Likelihood bootstrap values (500 replicates); GenBank accession numbers are given for each sequence.

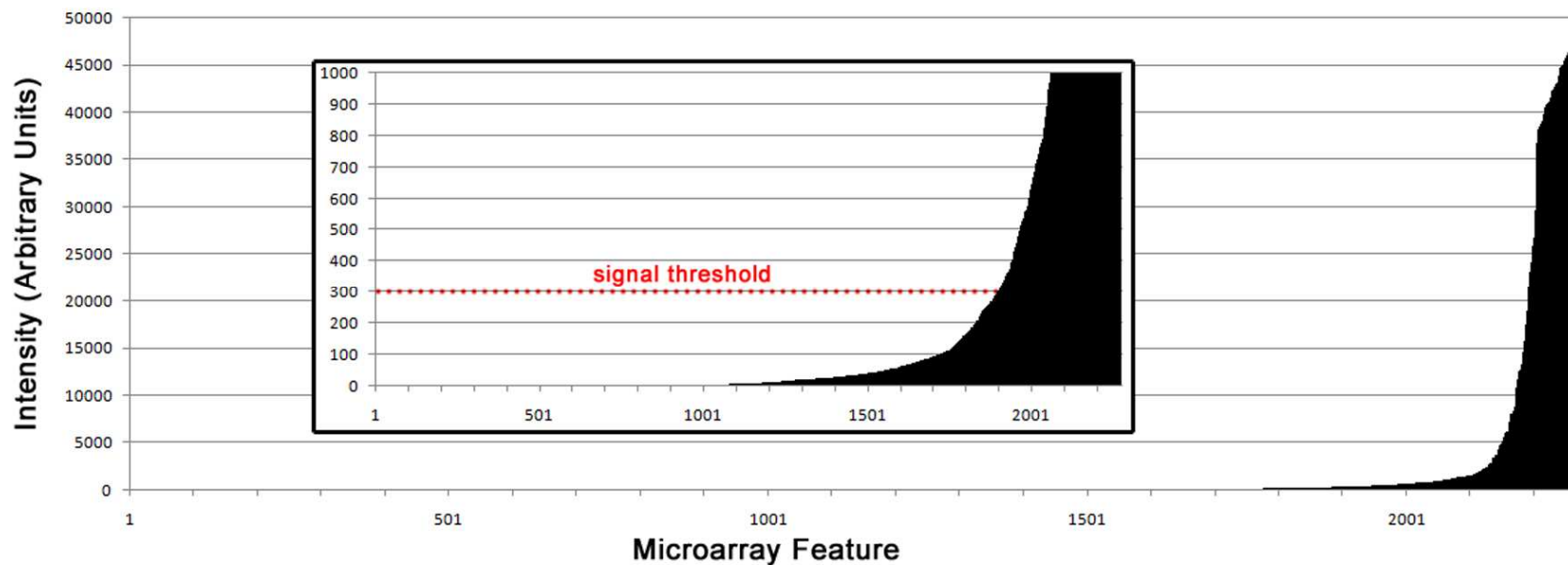


Fig. S15. Fluorescence intensity profile and threshold settings for a typical microarray hybridization. This profile shows the signal intensity distribution of a hybridization profile, for which the significance threshold was set to 300 units. The inset shows a magnification of the same profile to better visualize that the threshold was set in a region where the signal distribution became exponential.