Giant virus with a remarkable complement of genes infects marine zooplankton

Matthias G. Fischer[a], Michael J. Allen[b], William H. Wilson[c], and Curtis A. Suttle[a,d,e,1]

Departments of [a]Microbiology and Immunology, [d]Botany, and [e]Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; [b]Plymouth Marine Laboratory, Plymouth PL1 3DH, United Kingdom; and [c]Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, ME 04575-0475

Edited* by James L. Van Etten, University of Nebraska, Lincoln, NE, and approved October 4, 2010 (received for review June 2, 2010)

As major consumers of heterotrophic bacteria and phytoplankton, microzooplankton are a critical link in aquatic foodwebs. Here, we show that a major marine microflagellate grazer is infected by a giant virus, *Cafeteria roenbergensis* virus (CroV), which has the largest genome of any described marine virus (≈730 kb of double-stranded DNA). The central 618-kb coding part of this AT-rich genome contains 544 predicted protein-coding genes; putative early and late promoter motifs have been detected and assigned to 191 and 72 of them, respectively, and at least 274 genes were expressed during infection. The diverse coding potential of CroV includes predicted translation factors, DNA repair enzymes such as DNA mismatch repair protein MutS and two photolyases, multiple ubiquitin pathway components, four intein elements, and 22 tRNAs. Many genes including isoleucyl-tRNA synthetase, eIF-2γ, and an Elp3-like histone acetyltransferase are usually not found in viruses. We also discovered a 38-kb genomic region of putative bacterial origin, which encodes several predicted carbohydrate metabolizing enzymes, including an entire pathway for the biosynthesis of 3-deoxy-D-*manno*-octulosonate, a key component of the outer membrane in Gram-negative bacteria. Phylogenetic analysis indicates that CroV is a nucleocytoplasmic large DNA virus, with *Acanthamoeba polyphaga* mimivirus as its closest relative, although less than one-third of the genes of CroV have homologs in Mimivirus. CroV is a highly complex marine virus and the only virus studied in genetic detail that infects one of the major groups of predators in the oceans.

nucleocytoplasmic large DNA virus | horizontal gene transfer | viral evolution | DNA repair | 3-deoxy-D-manno-octulosonate

Predation by protistan grazers is a major pathway of carbon transfer and nutrient recycling in marine and freshwater systems (1); yet, viruses infecting phagotrophic protists in marine systems are largely unknown and completely unexplored genetically. The discovery of the giant *Acanthamoeba polyphaga* mimivirus in a freshwater amoeba, with its 1.2 million-base pair (bp) genome and 981 genes (2, 3), has sparked an intense debate about the biology and evolutionary origin of giant viruses. Whereas some researchers argue that giant viruses are "gene robbers" that have acquired their extensive gene collection by horizontal gene transfer (HGT) from cellular organisms (4–6), others favor the theory that these viruses date back to the emergence of eukaryotes and that most of their genes are viral in origin (7, 8). Recently, it has become evident that protists host the largest and most complex viruses known (9), that other giant viruses are likely widespread in oceans (10), and that some of these are pathogens of phytoplankton (11); yet, the only characterized giant viruses are those infecting species of *Acanthamoeba*. Ultimately, understanding the origin and evolution of giant viruses will be facilitated through the use of comparative genomics with other representative systems.

In this study, we used 454 pyrosequencing to sequence and de novo assemble the genome of a very large (300 nm capsid diameter) DNA virus, *Cafeteria roenbergensis* virus (CroV) strain BV-PW1, that was isolated from the coastal waters of Texas in the early 1990s (12). This lytic virus infects a marine heterotrophic flagellate, which is identical to *C. roenbergensis* strain VENT1 at the level of 18S rDNA. The host, which consumes bacteria and viruses (13), was originally misidentified as *Bodo* sp. (12). It is a 2-μm– to 6-μm–long bicosoecid heterokont phagotrophic flagellate (Stramenopiles) that is widespread in marine environments and is found in various habitats such as surface waters, deep sea sediments, and hydrothermal vents (14, 15). Populations of *C. roenbergensis* may be regulated by viruses in nature (16).

Results and Discussion

General Genome Features. The genome of CroV is a linear double-stranded DNA molecule with a size of ≈730 kb, making this the second largest described viral genome. We sequenced and assembled the 618-kb central part of the viral chromosome, which is flanked on both ends by large and highly repetitive regions (Fig. 1). These terminal regions could potentially serve as protective caps for the protein-coding part of the genome, akin to telomeres in eukaryotes. The CroV genome is AT-rich (77% A+T), which is reflected in the distribution of codons and in the overall amino acid (aa) composition. AT-rich codons are consistently preferred over GC-rich ones, with the four most frequent aa (Lys, Ile, Asn, and Leu) each representing ≈10% of the overall aa (*SI Appendix*, Fig. S1).

Using conservative annotation criteria (*SI Appendix*), we identified 544 putative protein-coding sequences (CDSs) in the 618-kb central region of the CroV genome, which had a coding density of 90.1%. The average CDS was 1,025 nucleotides (nt) in length, and coding capacities ranged from 47 to 3,337 aa. Applying a BLASTP E-value cutoff of 1*e*-05, 267 CDSs (49%) displayed similarity to sequences in GenBank and 134 CDSs (25%) could be assigned to one or more Clusters of Orthologous Groups of proteins (COGs, E <0.001) (*SI Appendix*, Fig. S2). CroV CDSs and their annotations are listed in Dataset S1. Based on the distribution of top BLASTP hits, approximately one-half of the CroV genes displayed similarities to proteins found in eukaryotes, bacteria, archaea, and other giant viruses (*SI Appendix*, Fig. S3). Twenty-two percent of CroV CDSs had their top BLASTP hit among eukaryotes, but in the absence of genomic information about *C. roenbergensis*, no statement can be made about potential gene transfer between CroV and its host. Although most CroV CDSs were of unknown function, 32% of CDSs could be assigned a putative function and they provide insights into the biology of this giant virus. Several of these enzymatic functions have not been reported to be encoded by any other virus (*SI Appendix*, Table S1).

Translation Genes. Viruses rely primarily on the protein translation apparatus of their hosts; it is therefore unusual to find viral genes

Author contributions: M.G.F., M.J.A., W.H.W., and C.A.S. designed research; M.G.F. and M.J.A. performed research; M.G.F. and M.J.A. analyzed data; and M.G.F. and C.A.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession nos. GU244497 (CroV genome) and GU249597 (partial 18S sequence from *C. roenbergensis* strain E4-10)]. The microarray data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE19051).

[1]To whom correspondence should be addressed. E-mail: csuttle@eos.ubc.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1007615107/-/DCSupplemental.

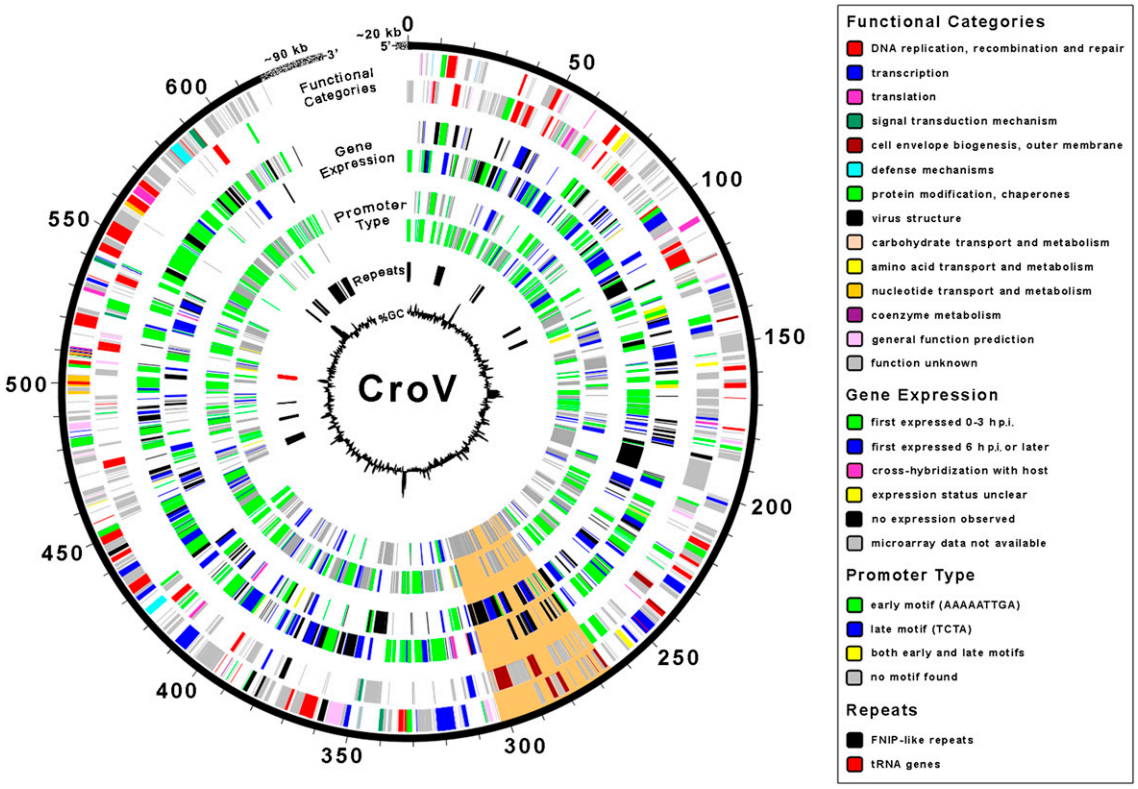


Fig. 1. Genome diagram of CroV. Genome coordinates are given in kbs. Nested circles from outermost to innermost correspond to (*i*) predicted CDSs on forward strand and (*ii*) reverse strand; (*iii*) expression data for CDSs on forward strand and (*iv*) reverse strand; (*v*) gene promoter type for CDSs on forward strand and (*vi*) reverse strand; (*vii*) location of repetitive DNA elements; (*viii*) GC content plotted relative to the genomic mean of 23.35% G+C. The speckled regions at the chromosome ends are not drawn to scale and indicate terminal repeats for which no sequence information is available. A 38-kb genomic segment of putative bacterial origin is shaded orange.

associated with protein synthesis. CroV encodes an isoleucyl-tRNA synthetase and putative homologs of eukaryotic translation initiation factors eIF-1, eIF-2α, eIF-2β/eIF-5, eIF-2γ, eIF-4AIII, eIF-4E, and eIF-5B. Using the transfer RNA gene prediction software tRNAScan-SE, we identified 22 tRNA genes, clustered in a 2.8-kb region around position 510,000 (Fig. 1 and *SI Appendix*, Table S2). We also found two putative tRNA-modifying enzymes in CroV, tRNA pseudouridine 5S synthase and $tRNA^{Ile}$ lysidine synthetase. These genes add to a rapidly growing number of virus-encoded protein translation components. Some tRNA genes are scattered among bacteriophages and eukaryotic viruses such as the phycodnaviruses (17, 18), and four tRNA synthetases along with several putative translation factors are found in Mimivirus (2). These findings imply that CroV and similarly complex viruses encode genes to modify and regulate the host translation system to their own advantage, which results in a "lifestyle" that is less dependent on host cell components than that of smaller viruses.

DNA Repair Genes. The ability to repair various kinds of DNA damage is well documented among large DNA viruses (19, 20). Given that the AT-rich genome of CroV is exposed to high solar irradiance in surface waters of the ocean and is therefore likely to suffer from DNA lesions such as pyrimidine dimers, it is not surprising that CroV encodes multiple DNA repair proteins. We found putative components of several DNA repair mechanisms, including a presumably complete base excision repair pathway with formamidopyrimidine DNA glycosylase, a family 1 apurinic/apyrimidinic (AP) endonuclease, a family X DNA polymerase, and an NAD-dependent DNA ligase. Further DNA repair proteins include DNA mismatch repair protein MutS, XPG endonuclease, a homolog of the alkylated DNA repair protein AlkB, and two DNA photolyases. Photolyases are classified into three major groups: Cyclobutane pyrimidine dimer (CPD) photolyases, (6-4) photolyases, and single-stranded DNA photolyases (ref. 21 and references therein). The CPD photolyases are further subdivided into class I and class II enzymes, the former being more prevalent in bacteria and the latter more frequent in eukaryotes. The gene product of crov115 is a predicted CPD class I photolyase and represents the first viral homolog in this class (*SI Appendix*, Figs. S4 and S5). The second CroV photolyase (crov149) does not belong to any of the established types of photolyases. Instead, it is related to a recently described group of photolyases/cryptochromes that are present in several bacterial phyla and the euryarchaeotes (21) (*SI Appendix*, Figs. S4 and S6). The only eukaryotic member in this group (*Paramecium tetraurelia*) is also the closest homolog to the CroV and Mimivirus sequences and may have acquired this gene by HGT from a giant virus (*SI Appendix*, Fig. S6).

Transcription Genes. Large DNA viruses typically carry hundreds of genes, including several that regulate gene expression. Among the predicted transcriptional genes in CroV are eight DNA-dependent RNA polymerase II subunits, at least six transcription factors involved in transcription initiation, elongation, and termination, a trifunctional mRNA capping enzyme, a poly(A) polymerase, and several helicases. The complex transcriptional machinery encoded by CroV suggests that viral gene transcription does not depend on host enzymes and likely occurs in the cytoplasm. Interestingly, CroV contains a CDS with high similarity to an ELP3-like histone acetyltransferase (HAT, COG1243, 2*e*-46), a gene previously not seen in viruses. In combination with other unidentified viral gene products, the CroV HAT may enable the virus to directly modulate the genome condensation state of the host and, thus, exert control over its transcriptional activity. Alternatively, this enzyme may be involved

in replication and packaging of the virus genome itself. Another unusual characteristic of CroV is the presence of three DNA topoisomerase (Topo) genes of types IA, IB, and IIA. TopoIA and TopoIIA are very similar to their counterparts in Mimivirus, and HGT events from bacteria (eventually via a eukaryotic phagotrophic host) have been proposed for these genes (22). CroV TopoIB is the first viral homolog of the eukaryotic subfamily, whereas the TopoIB encoded by Mimivirus falls within the bacterial group (*SI Appendix*, Fig. S7) and is functionally more similar to the poxvirus enzymes (23). Despite apparently different evolutionary trajectories, the presence of three Topo genes in CroV and Mimivirus suggests a crucial role for these enzymes in transcription, replication, or packaging of giant virus genomes.

Repetitive DNA and Ubiquitin Components. Approximately 5% of the genome (excluding the terminal regions) consisted of repetitive elements. The most prevalent was a 22-aa–long leucine-rich repeat similar to the FNIP/IP22 repeat (Pfam entry PF05725) that had >400 copies in the CroV genome and was present in at least 28 CDSs (Fig. 1 and Dataset S1). This repeat also occurs in Mimivirus and *Dictyostelium discoideum* (24). Whereas leucine-rich repeats are known to mediate protein–protein interactions in a variety of proteins with diverse functions (25), the role of these repeats in CroV is unknown. In Mimivirus, FNIP/IP22 repeat-containing genes also possess an N-terminal F-box domain, which mediates interaction with the ubiquitin (Ub) pathway (26). Ub signaling appears to be a general strategy used by nucleocytoplasmic large DNA viruses (NCLDVs) to counter host defenses, because multiple Ub-conjugating and Ub-hydrolyzing enzymes have been found in these viruses (26). Furthermore, it has been shown that orthopoxvirus replication requires a functional Ub-proteasome system (27). In CroV, we identified a small arsenal of genes encoding proteins predicted to function in the Ub pathway, including an E1 Ub-activating enzyme, six E2 Ub-conjugating enzymes, two deubiquinating enzymes, and one Ub gene. The specific means of how CroV and other giant viruses use Ub signaling to interact with their hosts remain to be determined.

CroV Harbors Four Inteins. No introns were detected in the genome, but four CDSs contained an intein, i.e., a self-splicing protein sequence inserted in highly conserved regions of a host protein (28). All four CroV inteins are part of NCLDV core genes that are thought to play a key role in DNA replication and transcription: DNA-dependent DNA polymerase B (PolB), TopoIIA, DNA-dependent RNA polymerase II subunit 2 (RPB2), and the large subunit of ribonucleotide reductase (RNR). Ten other inteins have been found in viruses infecting eukaryotes (29), including PolB inteins in Mimivirus (30), *Heterosigma akashiwo* virus (31), and *Chrysochromulina ericina* virus (32) as well as RNR inteins in four iridoviruses and the chlorella virus NY-2A (33). With the exception of a gene fragment from *Emiliania huxleyi* virus 163 (34), the CroV RPB2 intein constitutes the only viral report of an intein in RPB2. Finally, the CroV TopoIIA intein is a unique case of an intein in a DNA topoisomerase II gene, thus extending the known range of intein-containing genes. All four CroV inteins possess the conserved nucleophilic residues that are required for the standard splicing reaction [C/S at the N-terminal splice junction and N(C/S/T) at the C-terminal splice junction] (28) and are therefore probably capable of autocatalytic excision.

Microarray Analysis. A microarray experiment was undertaken to determine which CroV genes were unambiguously transcribed in infected cells and if there was a clear temporal pattern in the transcription of those genes. We detected viral transcripts in infected *C. roenbergensis* cells by fluorescently labeling mRNA isolated at different time points during the infection cycle, which lasted 12–18 h in *C. roenbergensis* strain E4-10. We then hybridized the labeled transcripts to glass slides spotted with oligonucleotide probes for 438 of the 544 predicted CroV genes (*SI Appendix*). Detectable levels of expression were found for 274 genes (63%), 152 genes (35%) were below the detection limit, 4 (1%) cross-hybridized with host mRNA isolated from uninfected cells, and 8 (2%) could not be assigned a clear on/off status (Fig. 1 and Dataset S1). Therefore, approximately one-half of the predicted genes and 63% of the genes we tested were expressed during infection under our laboratory conditions. This percentage is comparable with the observed expression of 65% of viral genes during infection of the marine phytoplankter *Emiliania huxleyi* by EhV-86 (35). However, recent gene expression studies in PBCV-1 and Mimivirus validated transcription for nearly all of their predicted genes (3, 36). It seems therefore likely that our microarray data underestimated the true extent of transcriptional activity in CroV. All of the previously mentioned translation-related genes in CroV, as well as most of the "virus-atypical" genes were expressed (Dataset S1), suggesting that these genes are functional. Although the microarray experiment was designed primarily to validate CroV gene predictions and cannot be exploited quantitatively, the data allowed us to recognize some general trends of CroV gene expression. Based on the time points at which transcripts were first detected, we could distinguish between an early and a late phase of CroV gene expression. The early phase lasted from 0 h to 3 h after infection (h p.i.) and affected 150 genes. The majority of DNA replication and transcription genes belonged to the early class. The late phase was characterized by genes that were first detected in the microarray at 6 h p.i. or later. The 124 genes in this class included all of the predicted structural components, such as the major and minor capsid proteins. Further and more extensive analysis of the CroV transcriptome may be able to refine this preliminary temporal classification.

Promoter Analysis. The intergenic regions had an average size of 71 ± 64 bp. We examined the 100-nt region upstream of the predicted start codons for possible promoter motifs by using MEME software (37). A perfectly conserved "AAAAATTGA" motif, flanked by AT-rich sequences, was found to precede 127 CroV CDSs (23%) (Fig. 2*A*). The MEME E-value for this motif was 9*e*-170. Allowing one mismatch per sequence at the less strongly conserved positions one to six of the AAAAATTGA motif increased the number of positive CDSs to 191 (35%). The majority of CDSs that displayed this motif in their immediate upstream region belonged to the "early" temporal category (Fig. 2*A*). We therefore classified this motif as an early gene promoter in CroV. Our results are in agreement with findings from Mimivirus, where a nearly identical early promoter motif (AAAATTGA) is associated with 45% of Mimivirus genes (3, 38). But, in contrast to Mimivirus, where the motif is found preferentially in the −50 to −110 region, the early promoter motif in CroV displayed a much narrower distribution, with a peak at position −40 relative to the predicted start codon (Fig. 2*B*).

We then searched for a possible late promoter motif starting with a representative set of six CDSs, all predicted to encode capsid components (*SI Appendix*). Five of the six genes exhibited the conserved tetramer "TCTA," flanked by AT-rich regions on either side, in their −11 to −20 region (Fig. 2). Based on this profile, we expanded the search to all CroV CDSs and identified 72 that were positive for the TCTA motif signature. A MEME search on the 30-nt upstream region of the 124 genes classified as "late" yielded a very similar motif (MEME E-value 5*e*-04; *SI Appendix*, Fig. S8). As shown in Fig. 2*A*, most CDSs with the TCTA promoter motif were first expressed at 6 h p.i. or later, supporting our conclusion that this sequence motif represents a promoter element for genes transcribed during the late phase of CroV infection. The CroV late promoter motif is unrelated to the putative late promoter motif identified in Mimivirus (3).

A Thirty-Eight–Kilobase Genomic Fragment Involved in Carbohydrate Metabolism. Upon examination of the CroV promoter distribution, we noticed that neither early not late promoter motifs were associated with CDSs located between the genomic positions 264,800 and 302,500 (Fig. 1). Of these 34 CDSs (crov242–crov275), 14 were most similar to bacterial proteins (*SI Appendix*, Table S3; BLASTP E-value <1*e*-05) and 7 of them are predicted to function in carbo-

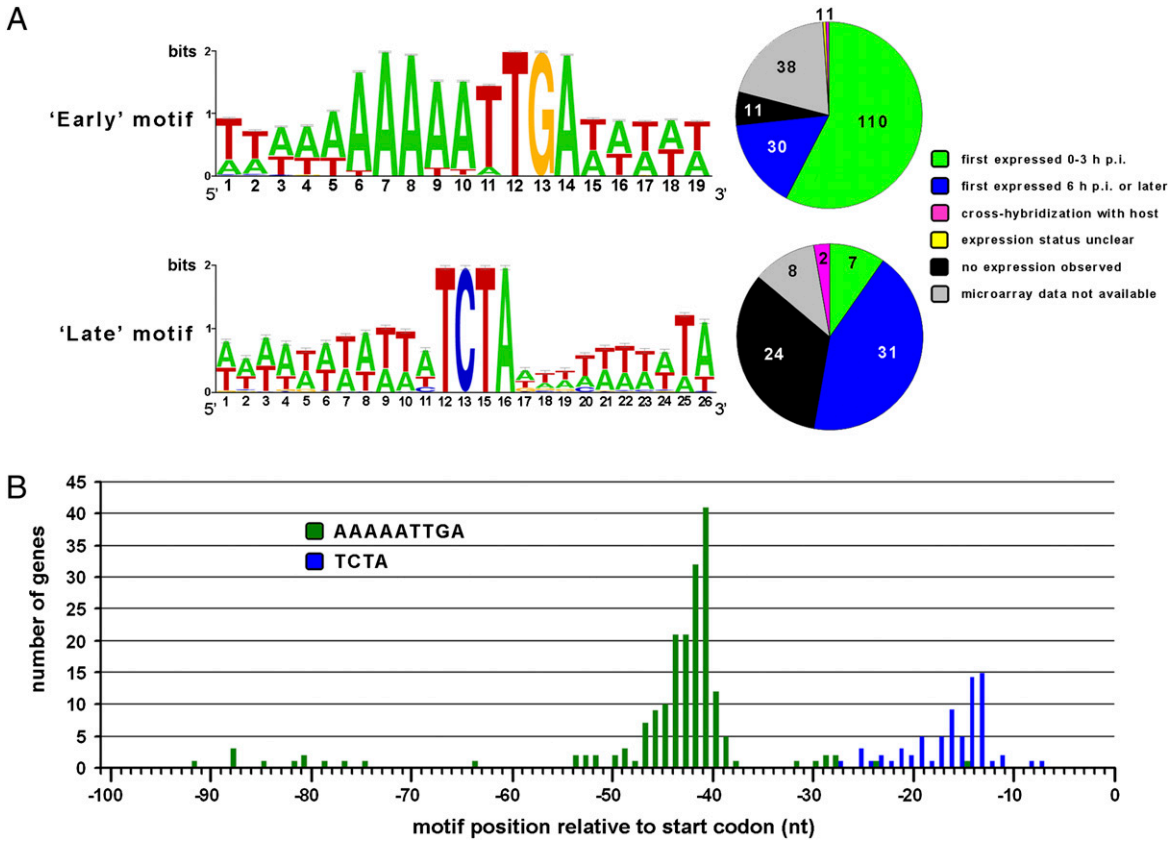


Fig. 2. Early and late gene promoter motifs in CroV. (*A*) Sequence logos depicting the consensus sequence for putative early (AAAAATTGA) and late (TCTA) promoter motifs. Pie charts show gene expression data for those CDSs that contained the respective motifs within their immediate 5′ upstream regions. The majority of CDSs associated with the AAAAATTGA motif were first seen expressed at 0–3 h p.i., whereas transcripts for most of the TCTA-associated CDSs were not detected until 6 h p.i. or later. (*B*) Positional distribution of the two motifs relative to the predicted start codon. A narrow distribution with a peak around position −40 is observed for the AAAAATTGA motif (n = 191). The TCTA motif (n = 72) occurs preferentially at position −13 to −21. The search for this motif was restricted to the upstream 30-nt region.

hydrate metabolism (Dataset S1). Among them, we identified enzymes for the biosynthesis of 3-deoxy-D-*manno*-octulosonate (KDO) (Fig. 3). In Gram-negative bacteria, KDO is an essential core component of the lipopolysaccharide layer, linking lipid A to polysaccharides (39). Biosynthesis of KDO, which is also found in the green alga *Chlorella* and the cell wall of higher plants, involves the three enzymes arabinose 5-phosphate isomerase (API), KDO 8-phosphate synthase (KDOPS), and KDO 8-phosphate phosphatase (KDOPase). A cytidylyltransferase (CMP-KDO synthetase, CKS) is then required to activate KDO for downstream reactions (Fig. 3*A*). We identified in crov265 a bifunctional KDOPase/API and in the N-terminal domain of crov267 a KDOPS homolog. The C-terminal domain of crov267 is a predicted dTDP-6-deoxy-L-hexose 3-*O*-methyltransferase, and crov266 encodes a predicted bifunctional *N*-acylneuraminate cytidylyltransferase (CMP-NeuAcS)/demethylmenaquinone methyltransferase (Fig. 3*B* and *SI Appendix*, Figs. S9–S12). Whether the cytidylyltransferase in crov266 is a functional CMP-NeuAcS and accordingly involved in sialic acid activation, as suggested by phylogenetic analysis (*SI Appendix*, Fig. S12), or rather a structurally related KDO-activating CKS, remains to be tested. The remaining CroV CDSs with functional annotation in this region are predicted glycosyltransferases and other sugar-modifying enzymes (Dataset S1). The presence of these genes and the finding that 10 of them were expressed (Dataset S1) suggests a role in viral glycoprotein biosynthesis and that the virion surface may be coated with KDO- or sialic acid-like glycoconjugates, which could be involved in virion-cell recognition. Given that the CDSs in this region lack the early/late promoter signals and have no homologs in Mimivirus, the 38-kb region must have been acquired after the CroV lineage split from the Mimivirus lineage. Because many of the CDSs in this region were most similar to bacterial genes (*SI Appendix*, Figs. S9 and S12 and Table S3), it is tempting to speculate that they

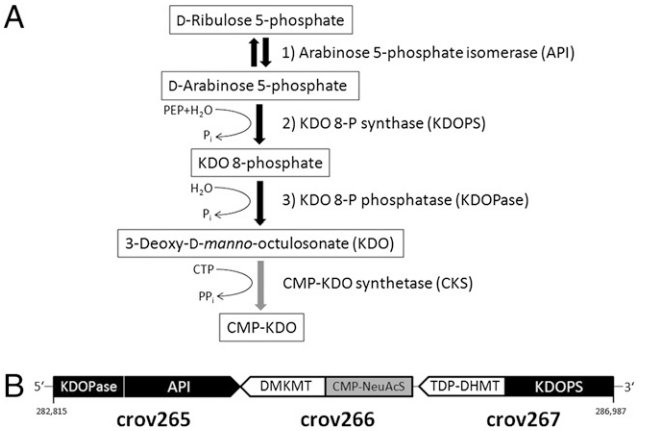


Fig. 3. The predicted KDO biosynthesis pathway in CroV. (*A*) Schematic of the three enzymatic steps that transform D-ribulose 5-phosphate into KDO. Activation of KDO to CMP-KDO is catalyzed by the cytidylyltransferase CKS. PEP, phosphoenolpyruvate. (*B*) Organization of the predicted KDO gene cluster (crov265–crov267) in the CroV genome. All three CDSs are predicted bifunctional enzymes. Genome coordinates are given. DMKMT, demethylmenaquinone methyltransferase; TDP-DHMT, dTDP-6-deoxy-L-hexose 3-*O*-methyltransferase; CMP-NeuAcS, *N*-acylneuraminate cytidylyltransferase.

may have been acquired from a bacterium, considering that CroV frequently encounters phagocytosed bacteria inside the host cytoplasm and encodes several enzymes that might catalyze an integration of foreign DNA (e.g., transposase crov356). Genomic islands of putative bacterial origin have been identified in other giant viruses such as phycodnaviruses and Mimivirus, but in contrast to CroV, these bacterial gene clusters tend to be located toward the ends of the linear viral chromosomes (40). However, given that the GC content of the 38-kb region is even lower than that of the rest of the CroV genome (19.4% vs. 23.6% G+C) and that some of these proteins occupy a phylogenetic position between bacterial and eukaryotic homologs (e.g., KDOPS and KDOPase; *SI Appendix*, Figs. S10 and S11), we cannot rule out alternative scenarios for the origin of this region.

Phylogenetic Relationship. Based on the presence and phylogenetic analysis of a set of core genes (*SI Appendix*, Fig. S13), CroV is an addition to the presumably monophyletic group of NCLDVs (2, 26, 41), which includes the families *Ascoviridae, Asfarviridae*, *Iridoviridae*, *Mimiviridae*, *Phycodnaviridae*, *Poxviridae*, and the newly discovered Marseillevirus (42). In a recent study by Yutin et al. (43), genes encoded by NCLDVs were categorized into groups that presumably evolved from a common ancestor and subsequently diversified in the various NCLDV families. Using this dataset of Nucleo-Cytoplasmic Virus Orthologous Genes (NCVOGs), we found that at least 172 CroV CDSs belonged to an existing NCVOG (Dataset S1). Thirty-two percent of CroV CDSs were significantly similar to a Mimivirus gene (any Mimivirus hit with a BLASTP E-value >1*e*-05) and 22 CroV CDSs had their only detectable GenBank homolog in Mimivirus. CroV therefore appears to be the closest known relative to Mimivirus, despite large differences in genome (730 kb vs. 1,181 kb) and capsid size (300 nm vs. 500 nm). The CroV–Mimivirus relationship was further corroborated by phylogenetic analysis of PolB, a commonly used marker gene to infer phylogenetic relationships among NCLDVs. Bayesian Inference analysis of PolB resulted in a strongly supported clade comprising the largest known viruses: Mimivirus, CroV, and three partially sequenced viruses infecting the marine microalgae *Phaeocystis pouchetii* (PpV), *Chrysochromulina ericina* (CeV), and *Pyramimonas orientalis* (PoV) (Fig. 4). These three algal viruses, for which only PolB and major capsid protein (MCP) sequences are available, also possess very large DNA genomes (485 kb, 510 kb, and 560 kb, respectively) and are proposed members of the family *Phycodnaviridae*, although a taxonomic revision of this tentative assignment has been proposed (10). Similarly, when the MCP was used to reconstruct the NCLDV phylogeny (*SI Appendix*, Fig. S14), these five viruses formed a monophyletic group that also included *Heterosigma akashiwo* virus, another large DNA virus that is assigned to the *Phycodnaviridae*. The topology of the NCLDV tree strongly suggests that the five largest viral genomes are more closely related to each other than to other NCLDV families and that they may have originated from a relatively recent ancestral virus that must have already been a bona fide NCLDV with a very large genome, probably encoding >150 proteins.

Conclusions

We present here the genetic analysis of a virus infecting a marine phagotroph. With a genome size larger than that of some cellular organisms, CroV is an example of an extraordinarily complex virus. It possesses a large number of predicted genes involved in DNA replication, transcription, translation, protein modification, and carbohydrate metabolism, indicating that CroV has a highly autonomous propagation strategy during infection.

The mechanisms by which such enormous virus genomes evolved have been much discussed (40, 44, 45). Most studies have focused on Mimivirus, because it represents the most extreme case of a giant virus and is the largest dataset available. The majority of Mimivirus genes have no cellular homologs and are presumably very ancient (46), up to one-third of its genes arose through gene and genome duplication (45), and <15% of Mimivirus genes may have been horizontally transferred from eukaryotes and bacteria (6). Our analysis of the CroV genome is consistent with this general picture of giant virus genome evolution. Gene duplication and lineage-specific expansion of the FNIP/IP22 repeat are two factors that clearly contributed to the enormous size of the CroV genome. Examples of duplicated genes are the paralogous groups of CDSs crov027–crov031 (contain FNIP/IP22 repeats), crov420–crov422 (unknown function), and some of the tRNA genes. A potential case of large-scale HGT from a bacterium is represented by the 38-kb genomic segment that differs in coding content and promoter regions from the rest of the viral genome. The remaining CDSs with cellular homologs are more difficult to categorize, because genes can be transferred from cells to viruses and vice

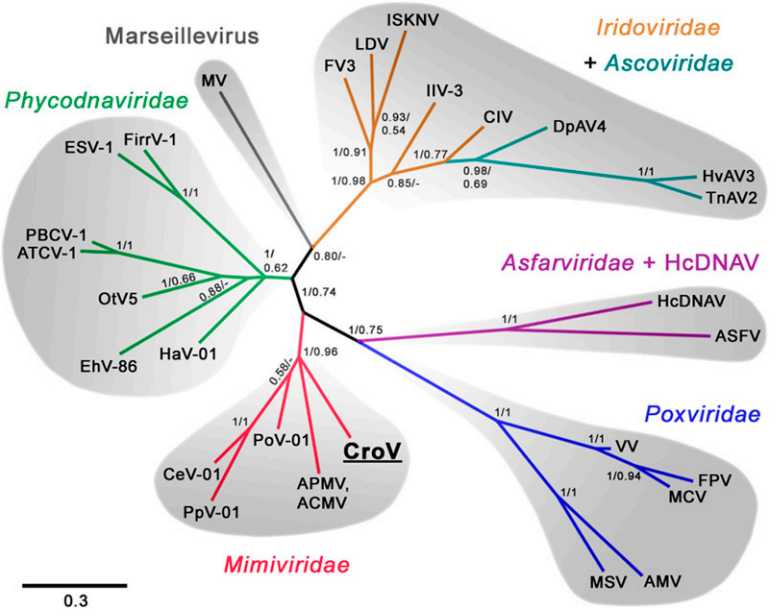


Fig. 4. Phylogenetic reconstruction of NCLDV members. The unrooted Bayesian Inference (BI) tree was generated from a 263-aa alignment of conserved regions of DNA polymerase B. Intein insertions were removed before alignment. Nodes are labeled with BI posterior probabilities and maximum likelihood bootstrap values (500 replicates). Abbreviations and accession numbers (GenBank unless stated otherwise) are as follows: ACMV, *Acanthamoeba castellanii* mamavirus, from ref. 43; AMV, *Amsacta moorei* entomopoxvirus, NP_064832; APMV, *A. polyphaga* mimivirus, YP_142676; ASFV, African swine fever virus, NP_042783; ATCV-1, *Acanthocystis turfacea* chlorella virus 1, YP_001427279; CeV-01, *C. ericina* virus 01, ABU23716; CIV, Chilo iridiscent virus, NP_149500; CroV, *C. roenbergensis* virus; DpAV4, *Diadromus pulchellus* ascovirus 4a, CAC19127; EhV-86, *E. huxleyi* virus 86, YP_293784; ESV-1, *Ectocarpus siliculosus* virus 1, NP_077578; FirrV-1, *Feldmannia irregularis* virus 1, AAR26842; FPV, Fowlpox virus, NP_039057; FV3, Frog virus 3, YP_031639; HaV-01, *H. akashiwo* virus 01, BAE06251; HcDNAV, *Heterocapsa circularisquama* DNA virus, DDBJ accession no. AB522601; HvAV3, *Heliothis virescens* ascovirus 3e, YP_001110854; IIV-3, Invertebrate iridiscent virus 3, YP_654692; ISKNV, Infectious spleen and kidney necrosis virus, NP_612241; LDV, Lymphocystis disease virus, YP_073706; MCV, Molluscum contagiosum virus, AAL40129; MSV, *Melanoplus sanguinipes* entomopoxvirus, NP_048107; MV, Marseillevirus, MAR_ORF329, GU071086; OtV5, *Ostreococcus tauri* virus 5, YP_001648316; PBCV-1, *P. bursarium* chlorella virus 1, NP_048532; PoV-01, *P. orientalis* virus 01, ABU23717; PpV-01, *P. pouchetti* virus 01, ABU23718; TnAV2, *Trichoplusia ni* ascovirus 2c, YP_803224; VV, Vaccinia virus, AAA98419.

versa. However, the majority of CroV CDSs show no significant similarity to any sequences in the public databases and their evolutionary origin remains hidden.

The array of "organismal" genes found in CroV further closes the overlap in metabolic coding capacity between large viruses and cellular life forms. This continued blurring of the distinction between what is considered living and nonliving adds to the ongoing debate about the puzzling evolutionary history of giant viruses (7, 8, 44). Moreover, the PolB gene of CroV has high similarity with those of other marine virus isolates, relatives of which appear to be widespread in the oceans (10), suggesting that CroV represents a major group of largely unknown but ecologically important marine viruses.

Materials and Methods

Flagellate Growth and Virus Purification. *C. roenbergensis* strain E4-10 was isolated from coastal waters near Yaquina Bay, OR, as described (13).

Cultures of *C. roenbergensis* were grown in f/2-enriched seawater medium supplemented with 0.01% (wt/vol) yeast extract to stimulate bacterial growth. The mixed assembly of bacteria in the cultures served as the food source for *C. roenbergensis.* Cultures were kept at room temperature (≈22 °C) in the dark. Typically, 1-L plastic Erlenmeyer flasks containing 250 mL of exponentially growing *C. roenbergensis* were infected at a cell density of 5×10^4 cells per mL by adding 100 μL (multiplicity of infection ≈0.5) of crude CroV-containing lysate. CroV purification is described in *SI Appendix*.

Genome Sequencing and Assembly. Phenol-chloroform extracted genomic DNA was sequenced by 454 pyrosequencing on GS 20 and GS FLX platforms. The two datasets were assembled individually and resulting contigs were analyzed with Sequencher (Gene Codes). Gap closing was achieved by a combination of multiplex PCR, bioinformatic prediction methods followed by PCR verification and sequencing, and a genomic shotgun library using the pSMART vector (Lucigen).

SI Appendix contains further details and experimental procedures on genome annotation and microarray analysis.

ACKNOWLEDGMENTS. We thank D. R. Garza, T. St. John, and A. S. Lang for help with the virus purification protocol during an early phase of the project, A. I. Culley and R. Adelshin for advice and discussion on DNA cloning and sequencing issues, A. M. Chan for assistance with flagellate culturing, and the staff at the Liverpool Microarray Facility for microarray fabrication. This work was supported by the Natural Science and Engineering Research Council of Canada Discovery Grants Program (to C.A.S.), the Tula Foundation through the Centre for Microbial Diversity and Evolution (C.A.S.), Natural Environment Research Council (NERC) Environmental Genomics Thematic Program Grants NE/A509332/1, NE/D001455/1WHW (to W.H.W.), and NERC SPG/MGF170 (to M.J.A.), and fellowships awarded by the Gottlieb Daimler- and Karl Benz-Foundation, Germany, and the University of British Columbia (to M.G.F.).

1. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* 3:537–546.
2. Raoult D, et al. (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306: 1344–1350.
3. Legendre M, et al. (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res* 20:664–674.
4. Filée J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol Biol* 8: 320.
5. Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* 7:306–311.
6. Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. *BMC Evol Biol* 8:12–21.
7. Koonin EV, Senkevich TG, Dolja VV (2006) The ancient Virus World and evolution of cells. *Biol Direct* 1:29.
8. Claverie JM (2006) Viruses take center stage in cellular evolution. *Genome Biol* 7:110.
9. Van Etten JL, Lane LC, Dunigan DD (2010) DNA viruses: The really big ones (giruses). *Annu Rev Microbiol* 64:83–99.
10. Monier A, et al. (2008) Marine mimivirus relatives are probably large algal viruses. *Virol J* 5:12.
11. Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G (2001) Isolation and characterization of two viruses with large genome size infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). *Virology* 290:272–280.
12. Garza DR, Suttle CA (1995) Large double-stranded DNA viruses which cause the lysis of a marine heterotrophic nanoflagellate (Bodo sp) occur in natural marine viral communities. *Aquat Microb Ecol* 9:203–210.
13. Gonzalez JM, Suttle CA (1993) Grazing by marine nanoflagellates on viruses and virus-sized particles: Ingestion and digestion. *Mar Ecol Prog Ser* 94:1–10.
14. Scheckenbach F, Wylezich C, Weitere M, Hausmann K, Arndt H (2005) Molecular identity of strains of heterotrophic flagellates isolated from surface waters and deep-sea sediments of the South Atlantic based on SSU rDNA. *Aquat Microb Ecol* 38: 239–247.
15. Atkins MS, Teske AP, Anderson OR (2000) A survey of flagellate diversity at four deep-sea hydrothermal vents in the Eastern Pacific Ocean using structural and molecular approaches. *J Eukaryot Microbiol* 47:400–411.
16. Massana R, del Campo J, Dinter C, Sommaruga R (2007) Crash of a population of the marine heterotrophic flagellate Cafeteria roenbergensis by viral infection. *Environ Microbiol* 9:2660–2669.
17. Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* 17:1486–1495.
18. Yamada T, Onimatsu H, Van Etten JL (2006) Chlorella viruses. *Adv Virus Res* 66: 293–336.
19. Srinivasan V, Schnitzlein WM, Tripathy DN (2001) Fowlpox virus encodes a novel DNA repair enzyme, CPD-photolyase, that restores infectivity of UV light-damaged virus. *J Virol* 75:1681–1688.
20. Furuta M, et al. (1997) Chlorella virus PBCV-1 encodes a homolog of the bacteriophage T4 UV damage repair gene denV. *Appl Environ Microbiol* 63:1551–1556.
21. Lucas-Lledó JI, Lynch M (2009) Evolution of mutation rates: Phylogenomic analysis of the photolyase/cryptochrome family. *Mol Biol Evol* 26:1143–1153.
22. Forterre P, Gribaldo S, Gadelle D, Serre MC (2007) Origin and evolution of DNA topoisomerases. *Biochimie* 89:427–446.
23. Benarroch D, Claverie JM, Raoult D, Shuman S (2006) Characterization of mimivirus DNA topoisomerase IB suggests horizontal gene transfer between eukaryal viruses and bacteria. *J Virol* 80:314–321.
24. O'Day DH, Suhre K, Myre MA, Chatterjee-Chakraborty M, Chavez SE (2006) Isolation, characterization, and bioinformatic analysis of calmodulin-binding protein cmbB reveals a novel tandem IP22 repeat common to many Dictyostelium and Mimivirus proteins. *Biochem Biophys Res Commun* 346:879–888.
25. Kobe B, Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* 11:725–732.
26. Iyer LM, Balaji S, Koonin EV, Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* 117:156–184.
27. Teale A, et al. (2009) Orthopoxviruses require a functional ubiquitin-proteasome system for productive replication. *J Virol* 83:2099–2108.
28. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E (2002) Inteins: Structure, function, and evolution. *Annu Rev Microbiol* 56:263–287.
29. Perler FB (2002) InBase: The Intein Database. *Nucleic Acids Res* 30:383–384.
30. Ogata H, Raoult D, Claverie JM (2005) A new example of viral intein in Mimivirus. *Virol J* 2:8.
31. Nagasaki K, Shirai Y, Tomaru Y, Nishida K, Pietrokovski S (2005) Algal viruses with distinct intraspecies host specificities include identical intein elements. *Appl Environ Microbiol* 71:3599–3607.
32. Larsen JB, Larsen A, Bratbak G, Sandaa RA (2008) Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* 74:3048–3057.
33. Fitzgerald LA, et al. (2007) Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect Chlorella NC64A. *Virology* 358:472–484.
34. Goodwin TJ, Butler MI, Poulter RT (2006) Multiple, non-allelic, intein-coding sequences in eukaryotic RNA polymerase genes. *BMC Biol* 4:38–54.
35. Wilson WH, et al. (2005) Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science* 309:1090–1092.
36. Yanai-Balser GM, et al. (2010) Microarray analysis of Paramecium bursaria chlorella virus 1 transcription. *J Virol* 84:532–542.
37. 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.
38. Suhre K, Audic S, Claverie JM (2005) Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. *Proc Natl Acad Sci USA* 102: 14689–14693.
39. Raetz CR (1990) Biochemistry of endotoxins. *Annu Rev Biochem* 59:129–170.
40. Filée J, Chandler M (2008) Convergent mechanisms of genome evolution of large and giant DNA viruses. *Res Microbiol* 159:325–331.
41. Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 75:11720–11734.
42. Boyer M, et al. (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* 106:21848–21853.
43. Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* 6:223.
44. Filée J, Siguier P, Chandler M (2007) I am what I eat and I eat what I am: Acquisition of bacterial genes by giant viruses. *Trends Genet* 23:10–15.
45. Suhre K (2005) Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. *J Virol* 79:14095–14101.
46. Ogata H, Claverie JM (2007) Unique genes in giant viruses: Regular substitution pattern and anomalously short size. *Genome Res* 17:1353–1361.

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

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 $1x10^3$ 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 $5x10^8$ 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℃ 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℃, 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 *Fsp*I, *Apa*I, and *Sac*II (all from NEB Canada), followed by PFGE separation of fragments. For each restriction analysis, $\approx 5x10^9$ 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℃ 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 1*e*-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 1*e*-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://mobyle.pasteur.fr/cgi-bin/MobylePortal/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 $1x10^5$ 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 RNAlater solution (Qiagen, Mississauga, ON, Canada) and stored at -80˚C until further use. RNA extraction was performed using an RNAeasy 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 µl 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 RNAeasy 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 NH_4Ac 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 42ºC for 2 hours followed by 95ºC 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 37ºC 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 95ºC and 24 cycles of 30 sec at 95ºC, 30 sec at 55ºC, 3 min at 72ºC. 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 µl transcription enzyme blend. The reaction was incubated for 16 hours at 37ºC 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 42ºC 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 µl 20x SSC, 1.2 µl 10% SDS, and the labeled cRNA samples were combined in a total volume of 60 µl and prewarmed to 60ºC. 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 60ºC 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 strain18S 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 nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* 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. Shackelton 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 2α, SUI2 homolog	crov162	Translation
Eukaryotic translation initiation factor 2γ	crov479	Translation
Eukaryotic translation initiation factor 5B	crov113	Translation
Isoleucyl-tRNA synthetase	crov505	Translation
tRNA pseudouridine 5S synthase	crov071	Translation
Bifunctional 3-deoxy-D-*manno*-2-octulosonate 8-P phosphatase / arabinose 5-phosphate isomerase	crov265	Lipopolysaccharide biosynthesis
Bifunctional N-acylneuraminate cytidylyltransferase / demethylmenaquinone methyltransferase	crov266	Lipopolysaccharide biosynthesis
Bifunctional 3-deoxy-D-*manno*-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

Fischer et al. doi:10.1073/pnas.1007615107

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

Fischer et al. doi:10.1073/pnas.1007615107

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

Fischer et al. doi:10.1073/pnas.1007615107

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

Fischer et al. doi:10.1073/pnas.1007615107

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

Fischer et al. doi:10.1073/pnas.1007615107

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

Fischer et al. doi:10.1073/pnas.1007615107

4. Supporting Figures

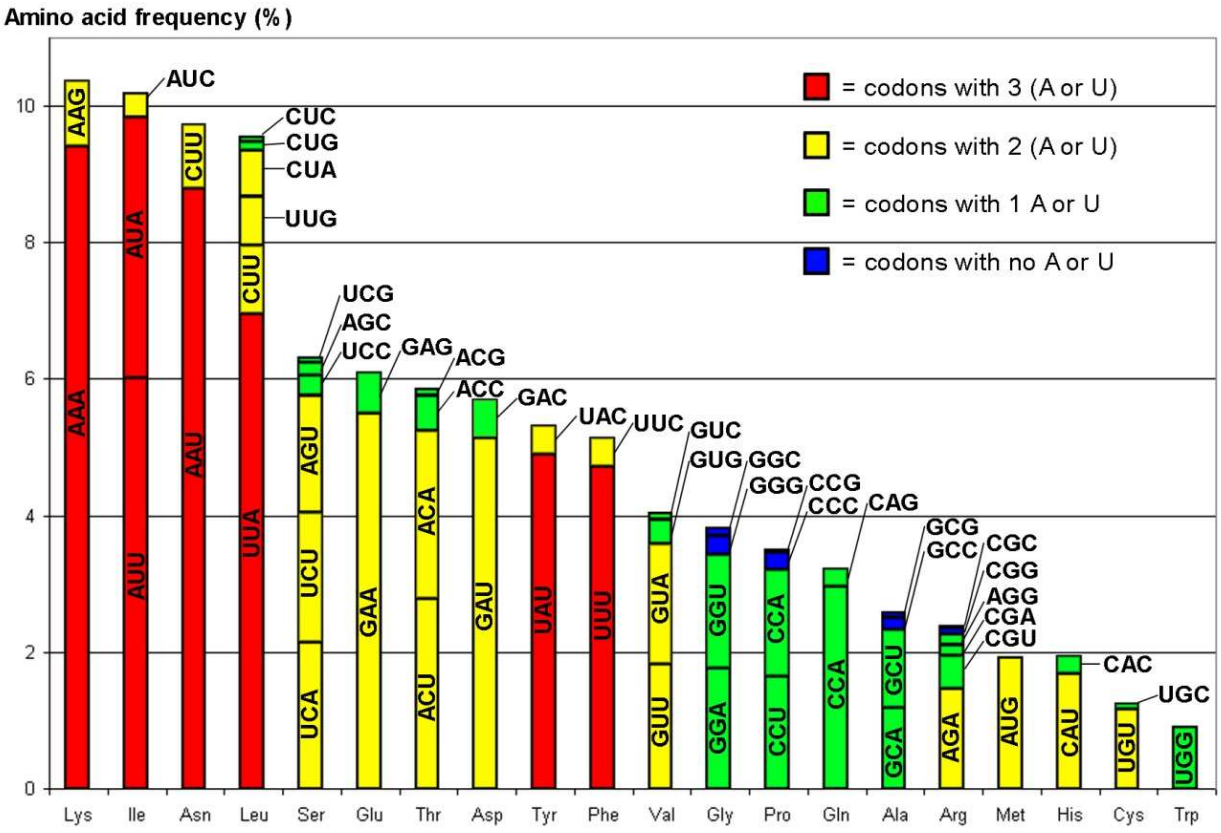


Fig. S1. Codon usage in CroV. This codon analysis is based on 185,006 codons in 544 CDSs (stop codons excluded). Codons consisting 100% of (A or U) are colored in red, 67% (A or U) in yellow, 33% (A or U) in green, 0% (A or U) in blue. The height of each codon column represents the overall frequency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

Fischer et al. doi:10.1073/pnas.1007615107

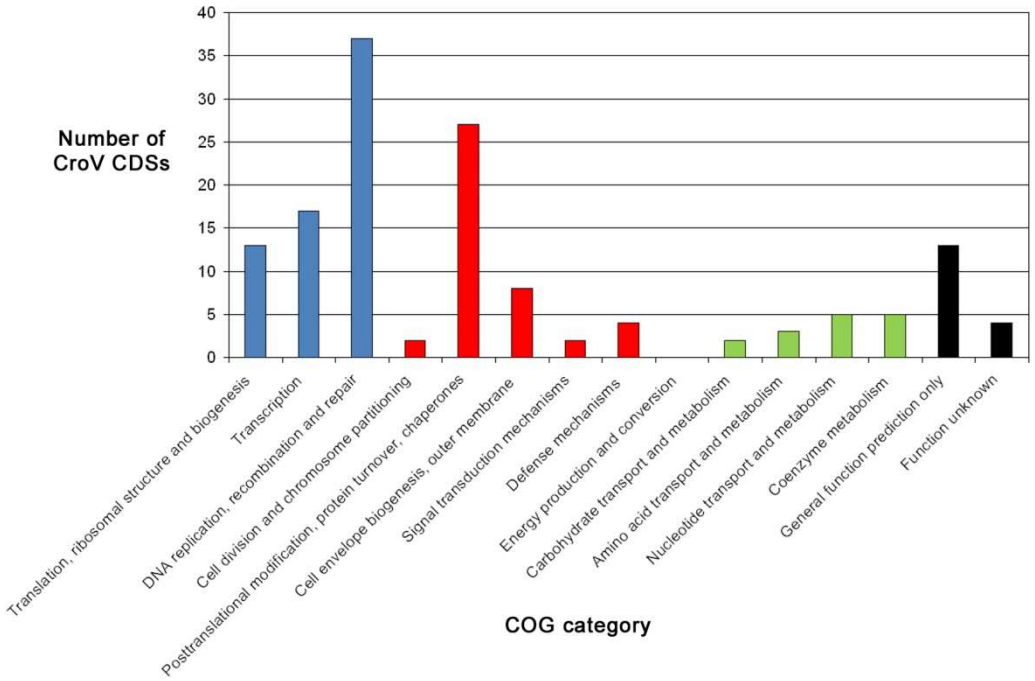


Fig. S2. Functional categories of Clusters of Orthologous Groups of proteins (COGs) identified in CroV. Color code: blue, information storage and processing; red, cellular processes; green, metabolism-related categories; black, poorly characterized categories.

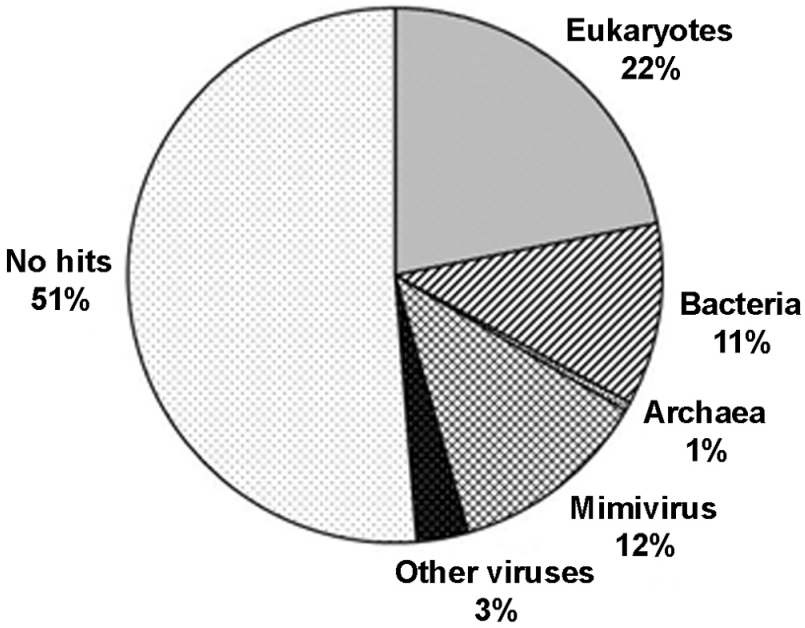


Fig. S3. Distribution of top BLASTP hits for CroV CDSs.
All 544 CroV CDSs were queried against the NCBI non-redundant database and categorized according to the domain affiliation of their top BLASTP hit. The E-value cutoff for this analysis was 1e-05.

Fischer et al. doi:10.1073/pnas.1007615107

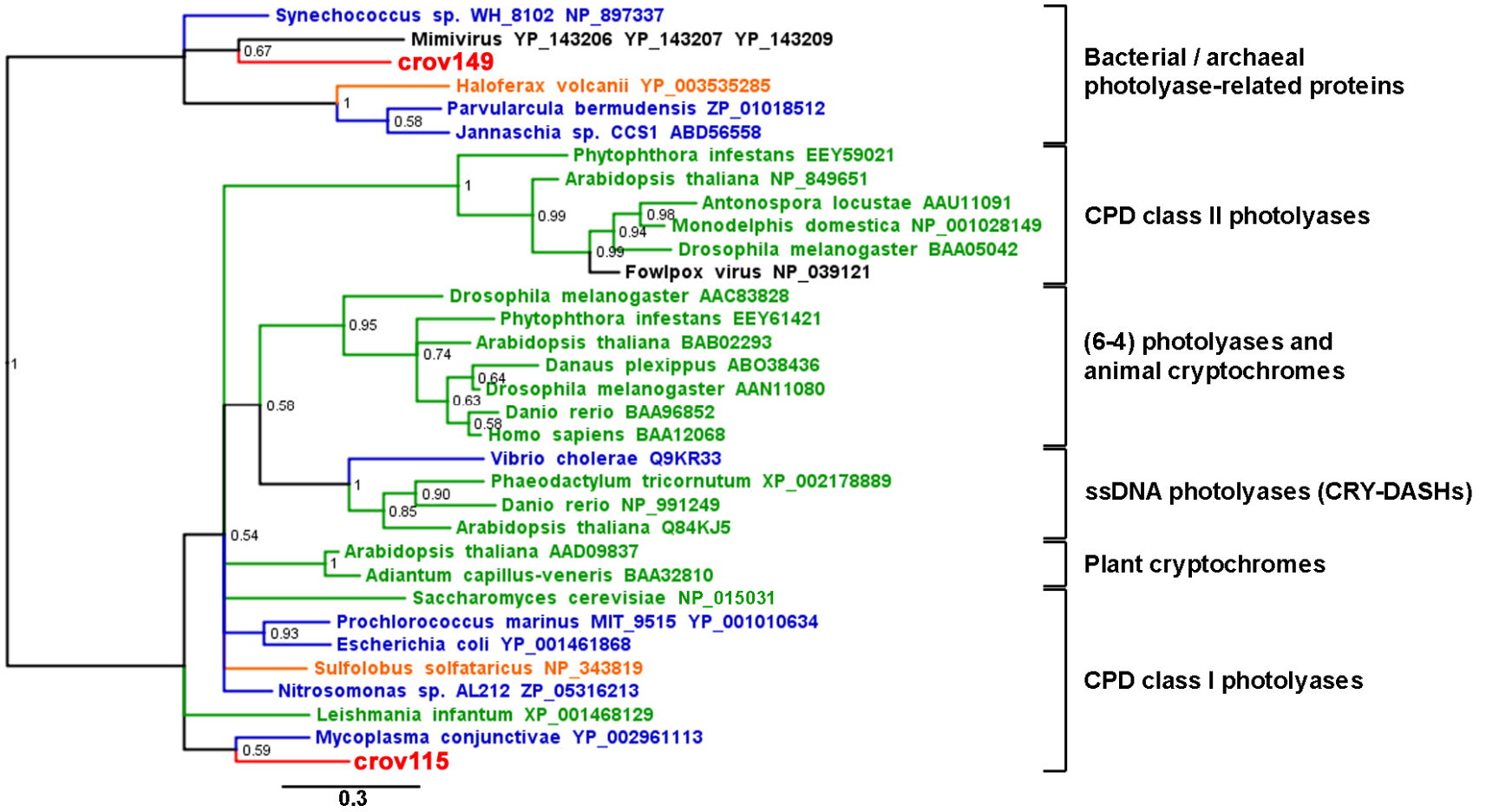


Fig. S4. Phylogenetic analysis of the photolyase/chryptochrome family. The unrooted Bayesian Inference tree of photolyases and chryptochromes is based on 73 conserved sites. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with posterior probabilities 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.

Fischer et al. doi:10.1073/pnas.1007615107

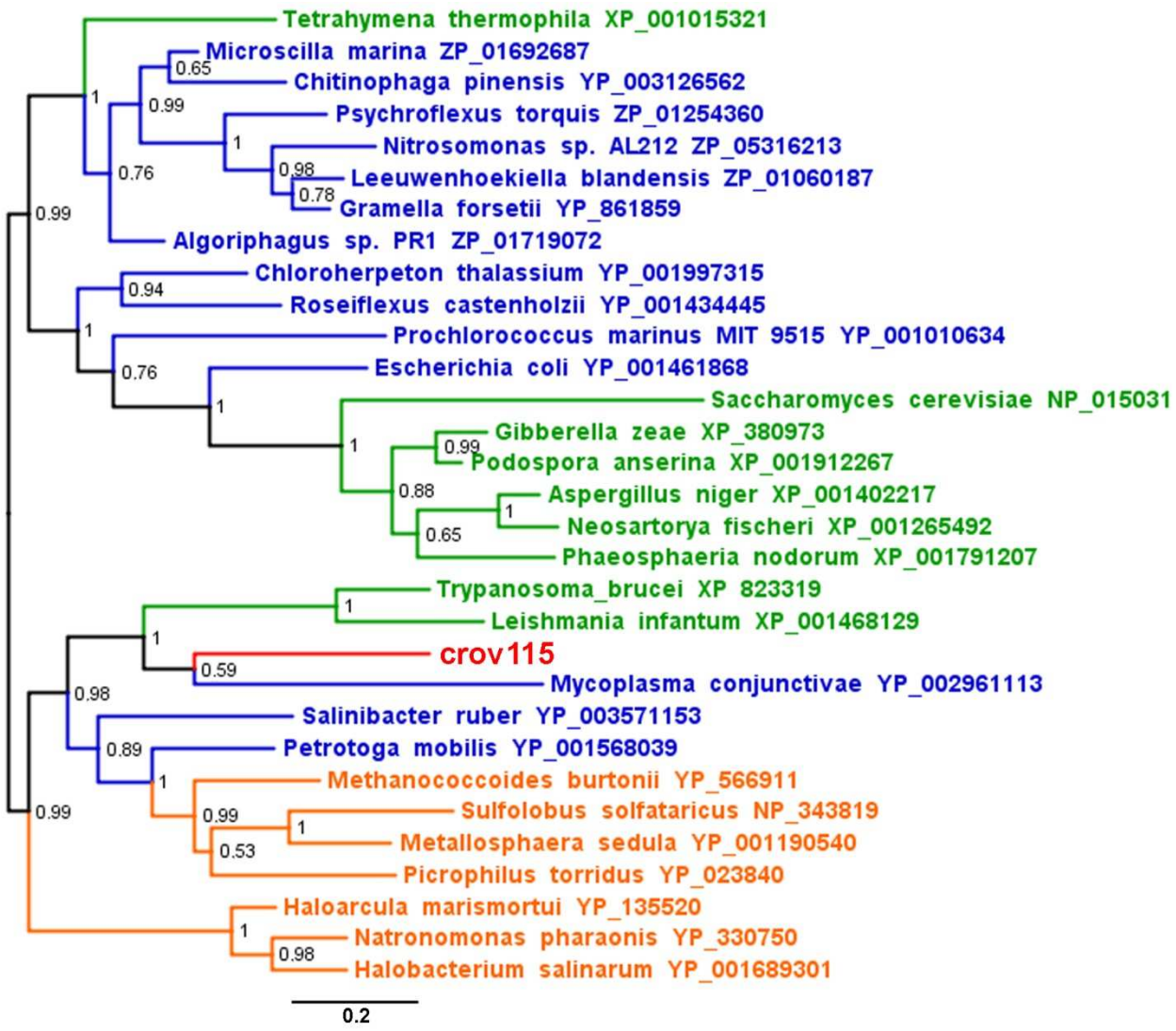


Fig. S5. Phylogenetic analysis of CPD class I photolyases. The unrooted Bayesian Inference tree is based on 189 conserved sites of CPD class I photolyases related to crov115. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with posterior probabilities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

Fischer et al. doi:10.1073/pnas.1007615107

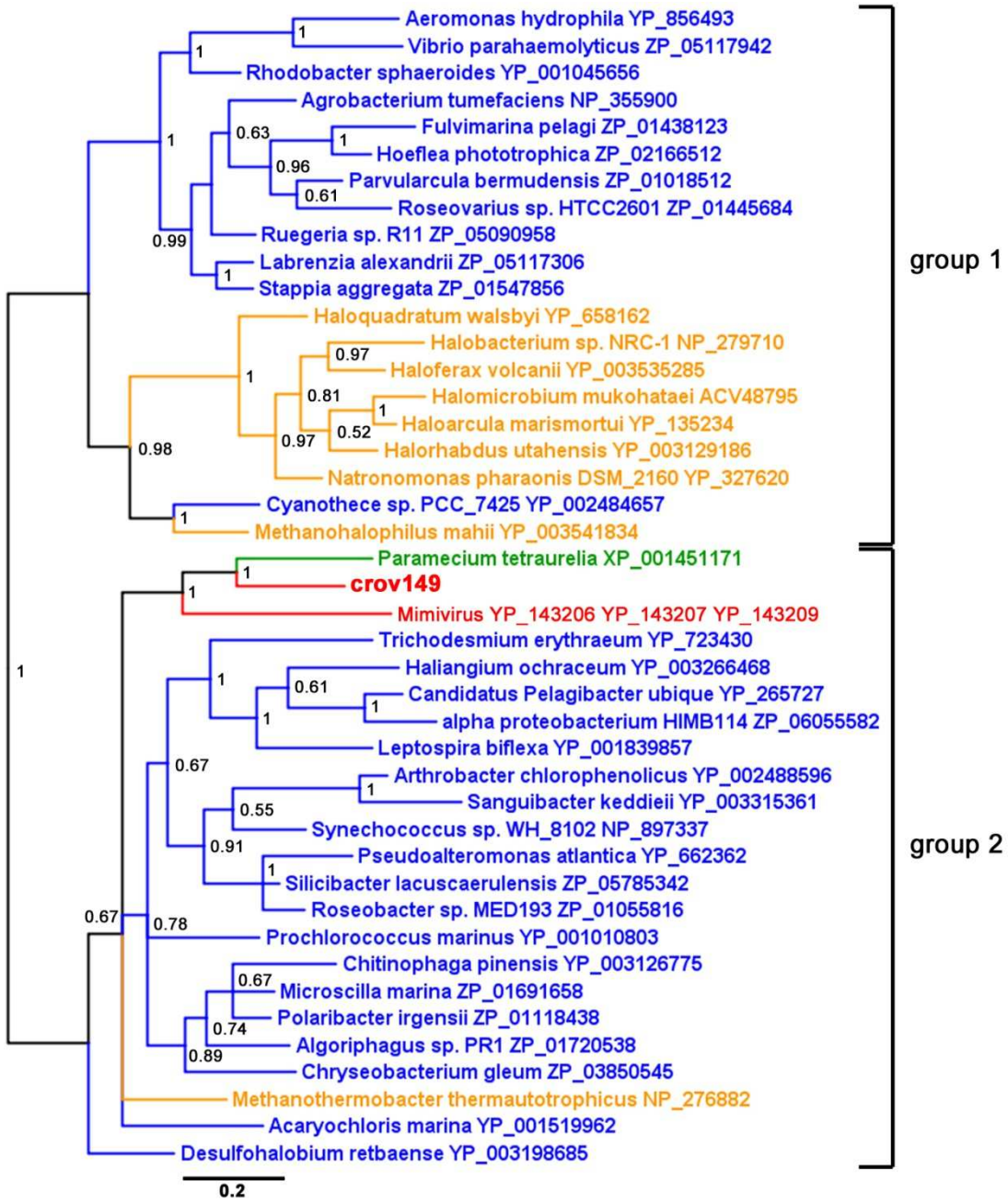


Fig. S6. Phylogenetic position of the predicted DNA photolyase crov149. The unrooted Bayesian Inference tree is based on a 157-aa alignment of DNA photolyases related to crov149. Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Two main groups can be differentiated, 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.

Fischer et al. doi:10.1073/pnas.1007615107

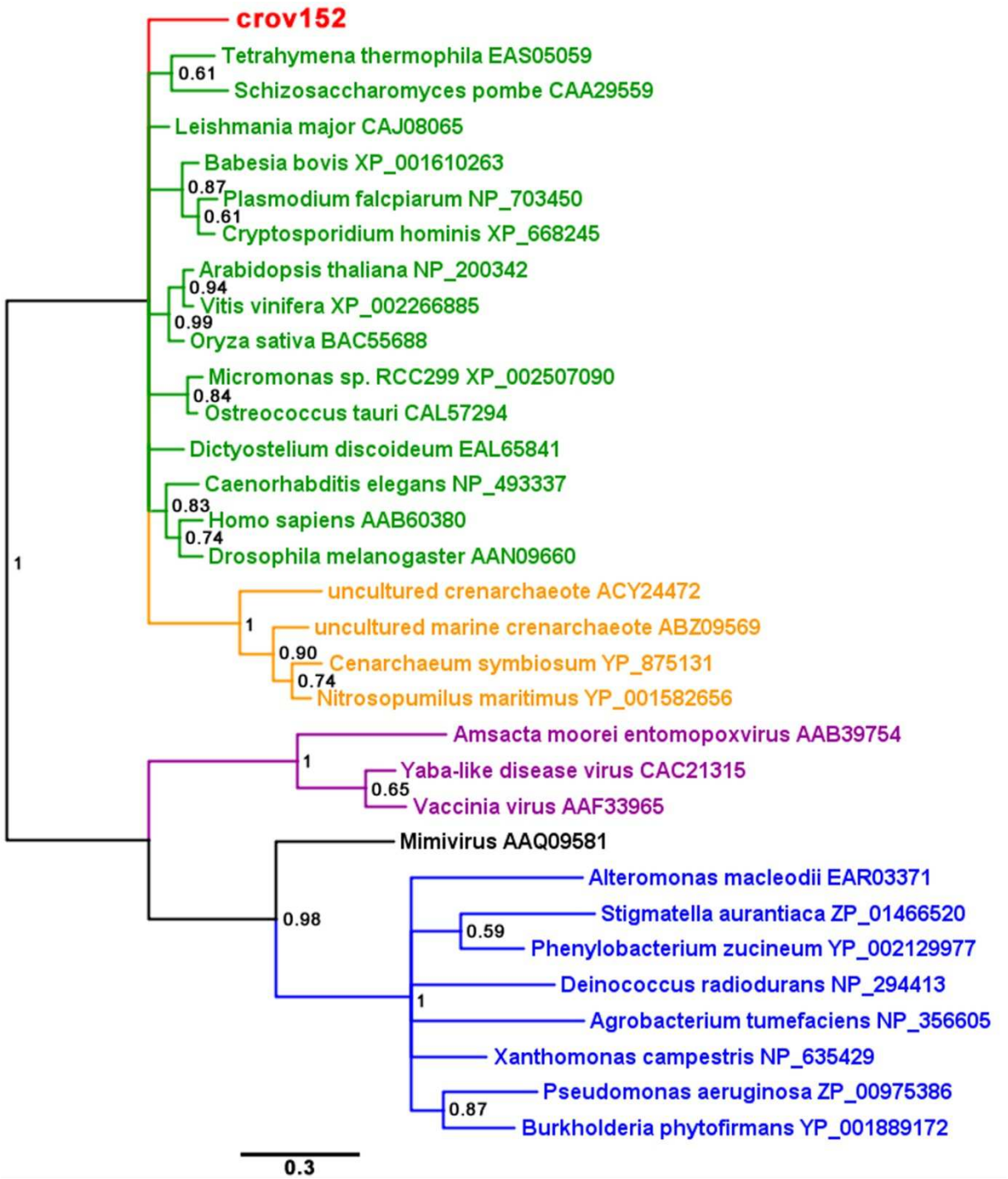


Fig. S7. Phylogenetic position of CroV DNA topoisomerase IB. The unrooted Bayesian Inference tree is based on a 66-aa alignment of DNA toposiomerases of type IB. Sequences are colored orange for archaea, blue for bacteria, green for eukaryotes and purple for poxviruses. Nodes are labeled with posterior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

Fischer et al. doi:10.1073/pnas.1007615107

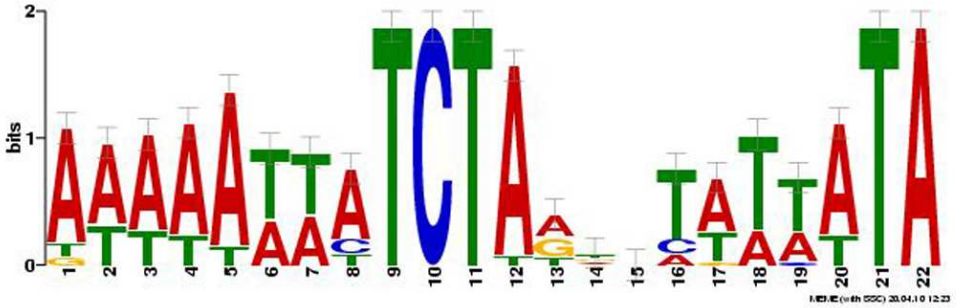


Fig. S8. MEME sequence logo of the CroV late promoter motif.

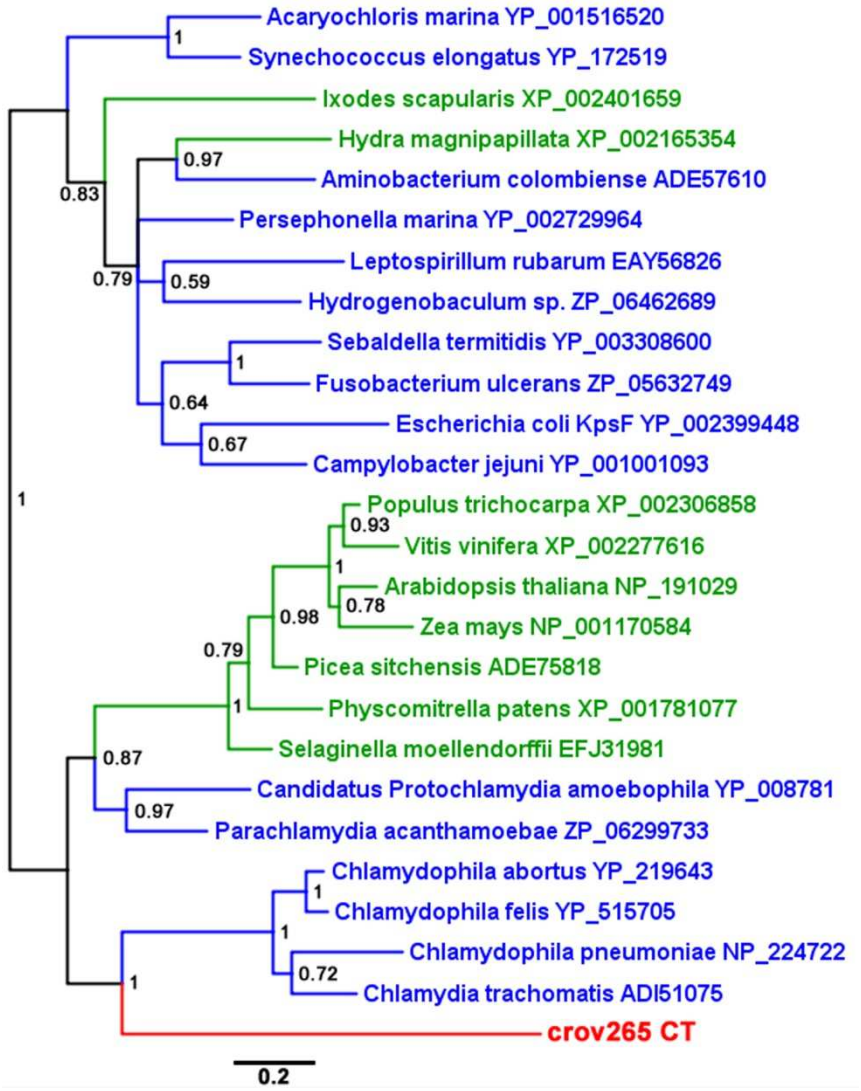


Fig. S9. Phylogenetic tree of API. The unrooted Bayesian Inference tree is based on a 228-aa alignment of arabinose-5-phosphate isomerases (API). Sequences are colored blue for bacteria and green for eukaryotes. Nodes are labeled with support values and GenBank accession numbers are given for each sequence.

Fischer et al. doi:10.1073/pnas.1007615107

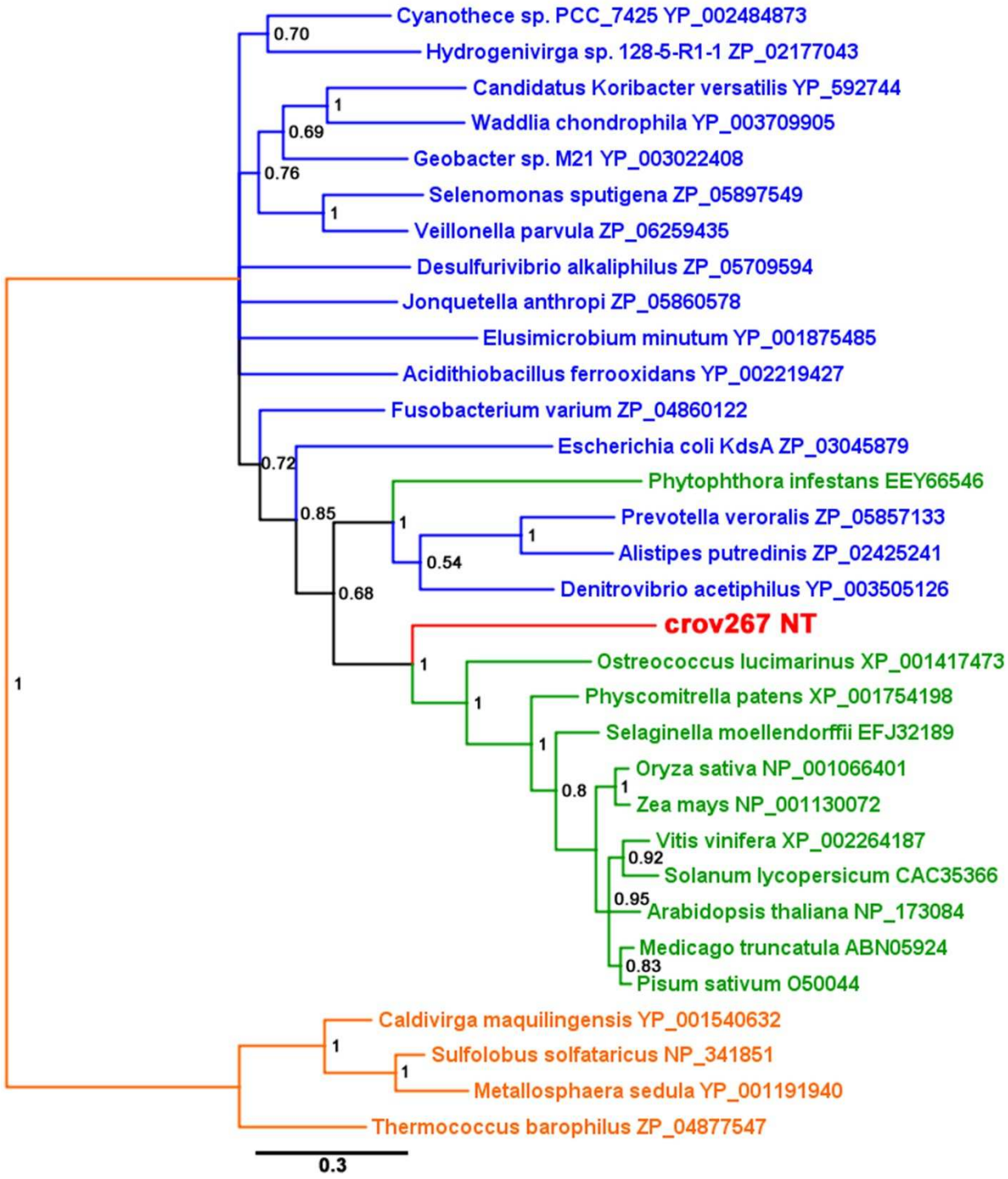


Fig. S10. Phylogenetic tree of KDO 8-P synthases. The unrooted Bayesian Inference tree is based on a 208-aa alignment of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthases (KDOPS). Sequences are colored orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labeled with support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

Fischer et al. doi:10.1073/pnas.1007615107

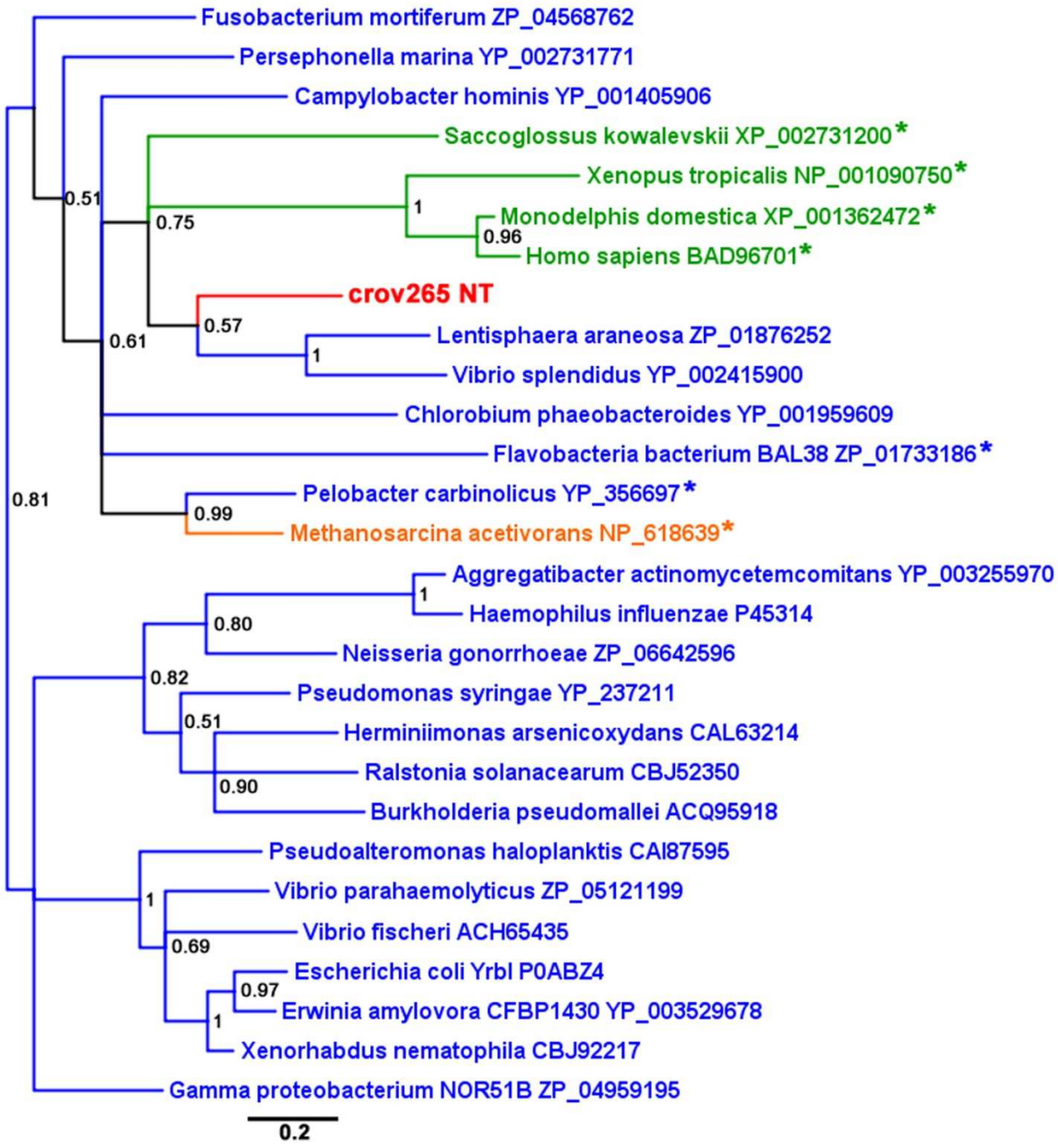


Fig. S11. Phylogenetic analysis of KDO 8-P phosphatases. The unrooted Bayesian Inference tree is based on a 134-aa alignment of 3-deoxy-D-*manno*-octulosonate 8-phosphate phosphatases (KDOPase), which belong to the haloacid dehalogenase-like (HAD) superfamily. Sequences marked with an asterisk are bifunctional enzymes where the C-terminal HAD domain is preceded by a N-acetylneuraminate cytidylyltransferase 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.

Fischer et al. doi:10.1073/pnas.1007615107

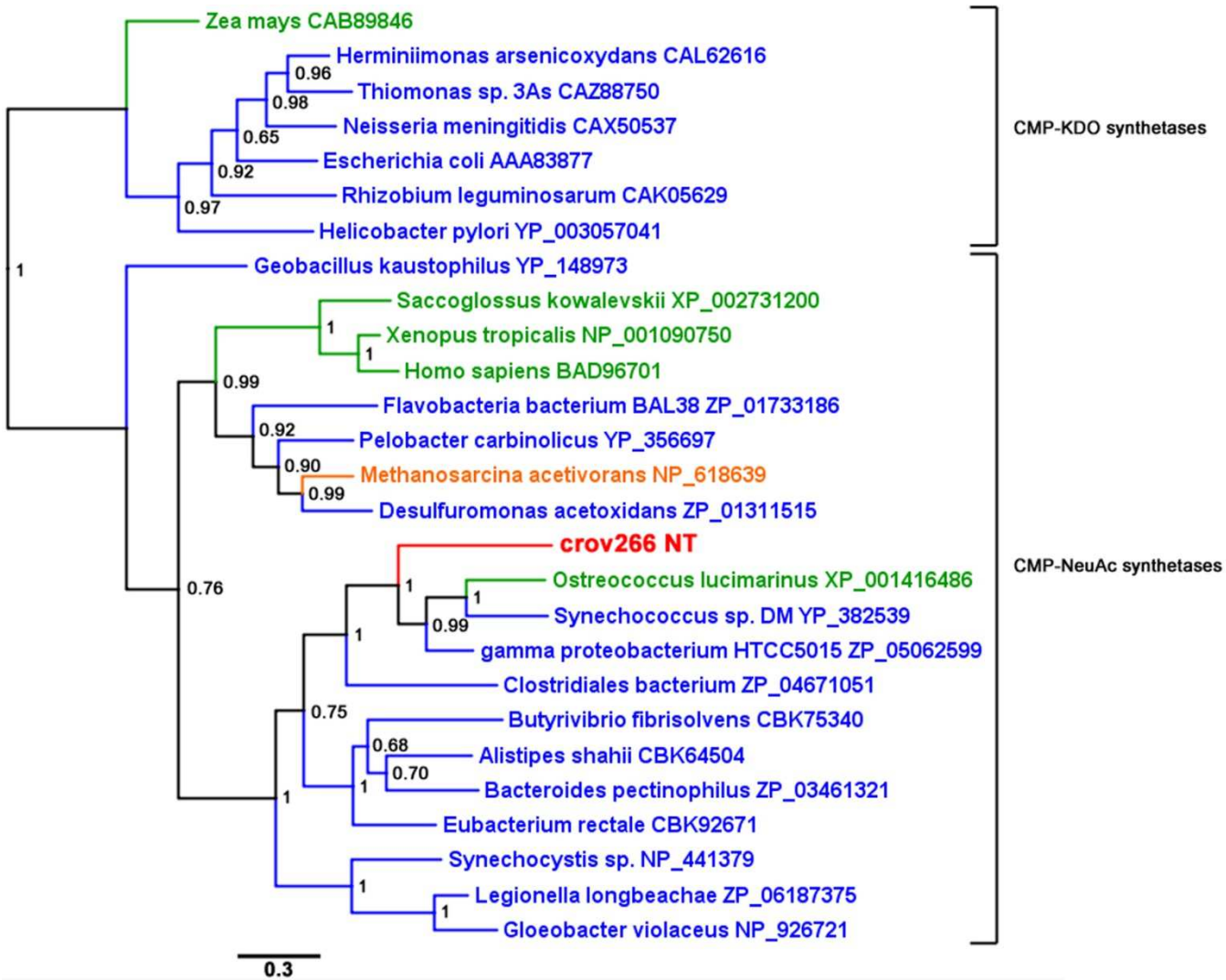
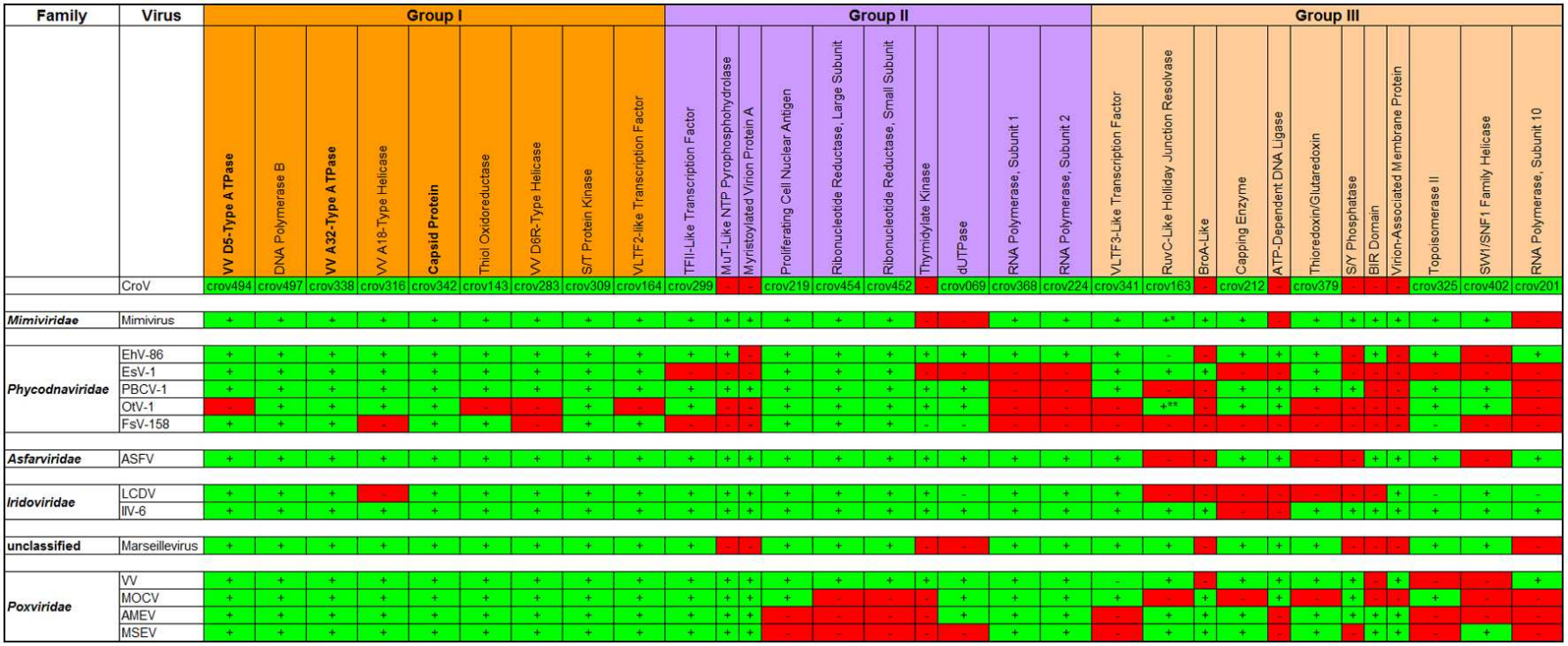


Fig. S12. Phylogenetic tree of two types of cytidylyltransferases. The unrooted Bayesian Inference tree is based on a 185-aa alignment of N-acylneuraminate cytidylyltransferases (CMP-NeuAc synthetases) and 3-deoxy- D-*manno* -octulosonate cytidylyltransferases (CMP-KDO synthetases). Sequences are colored 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.

Fischer et al. doi:10.1073/pnas.1007615107



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	VLTF2-like Transcription Factor	TFII-Like Transcription Factor	MuT-Like NTP Pyrophosphohydrolase	Myristoylated Virion 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	VLTF3-Like Transcription Factor	RuvC-Like Holliday Junction Resolvase	BroA-Like	Capping Enzyme	ATP-Dependent DNA Ligase	Thioredoxin/Glutaredoxin	S/Y Phosphatase	BIR Domain	Virion-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

Fischer et al. doi:10.1073/pnas.1007615107

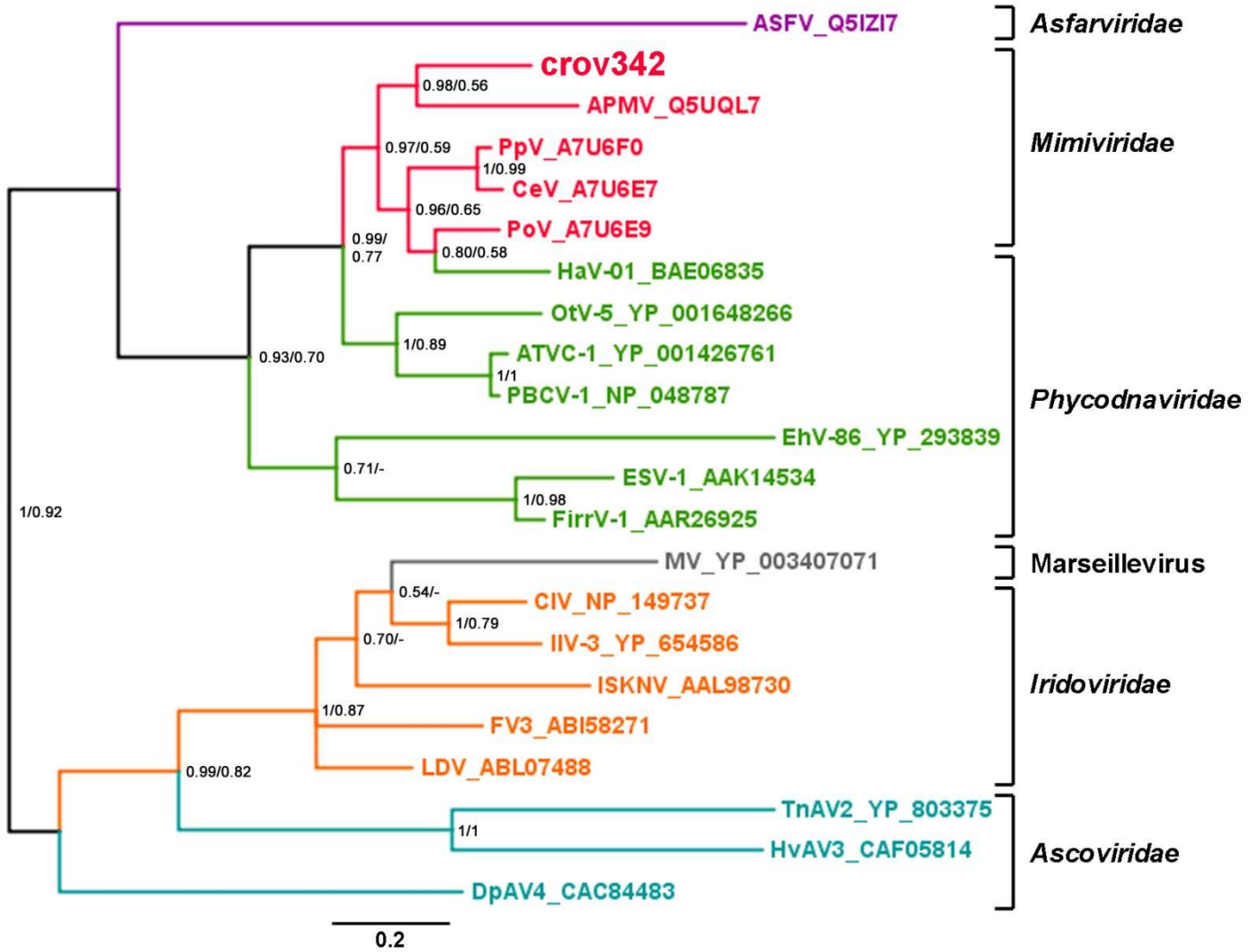


Fig. S14. Phylogenetic analysis of the NCLDV major capsid protein. The unrooted Bayesian Inference tree is based on a 169-aa alignment. Color coding and abbreviations are the same as in Fig. 4. Not included in the tree are the poxviruses, as their capsid proteins are too divergent from those of other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

Fischer et al. doi:10.1073/pnas.1007615107

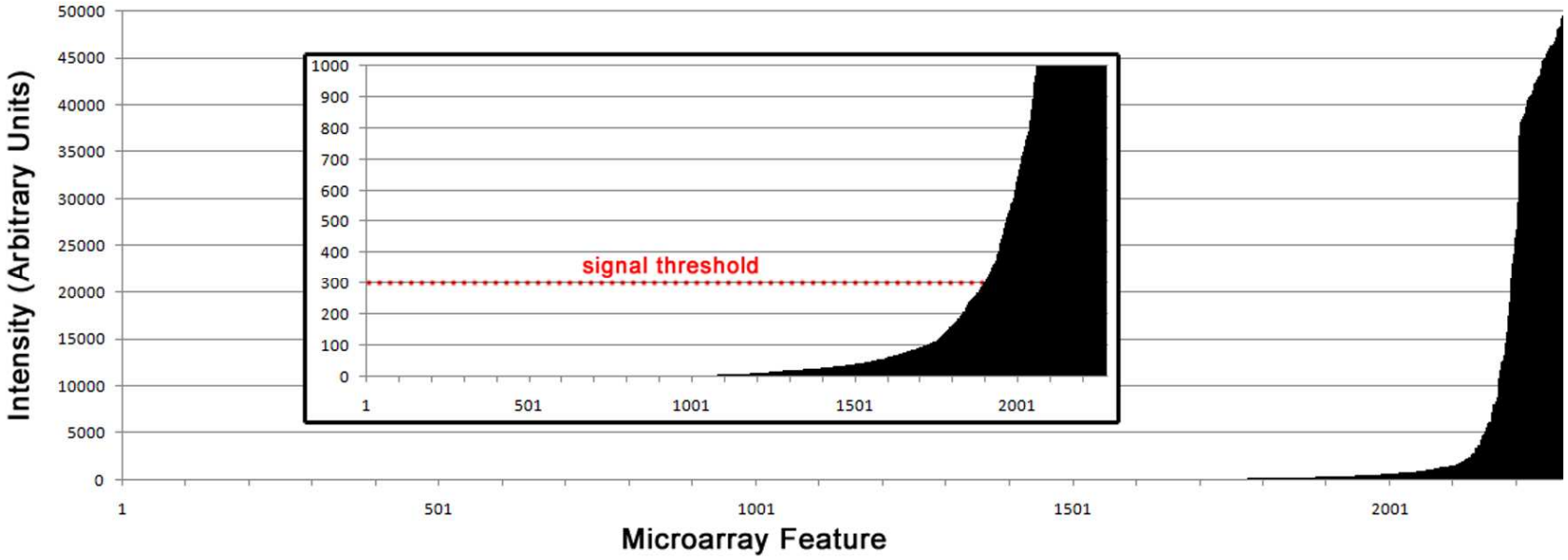


Fig. S15. Fluorescence intensity profile and threshold settings for a typical microarray hybridization. This profile shows the signal intensity distribution of a hybridization profile, for which the significance threshold was set to 300 units. The inset shows a magnification of the same profile to better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

Fischer et al. doi:10.1073/pnas.1007615107