Virology 466-467 (2014) 82–94

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

The virion of *Cafeteria roenbergensis* virus (CroV) contains a complex suite of proteins for transcription and DNA repair

Matthias G. Fischer [a,*], Isabelle Kelly [b,1], Leonard J. Foster [b], Curtis A. Suttle [a,c,d]

[a] Department of Microbiology & Immunology, University of British Columbia, Vancouver, Canada V6T 1Z4
[b] Department of Biochemistry & Molecular Biology and Centre for High-Throughput Biology, University of British Columbia, Vancouver, Canada V6T 1Z4
[c] Department of Botany, University of British Columbia, Vancouver, Canada V6T 1Z4
[d] Department of Earth & Ocean Sciences, University of British Columbia, Vancouver, Canada V6T 1Z4

ARTICLE INFO

Available online 25 June 2014

Keywords:
Cafeteria roenbergensis virus
CroV
NCLDV
Megavirales
Mimivirus
Giant virus
Virophage
Virion
Proteome
LC–MS/MS

ABSTRACT

Cafeteria roenbergensis virus (CroV) is a giant virus of the *Mimiviridae* family that infects the marine phagotrophic flagellate *C. roenbergensis*. CroV possesses a DNA genome of ~730 kilobase pairs that is predicted to encode 544 proteins. We analyzed the protein composition of purified CroV particles by liquid chromatography–tandem mass spectrometry (LC–MS/MS) and identified 141 virion-associated CroV proteins and 60 host proteins. Data are available via ProteomeXchange with identifier PXD000993. Predicted functions could be assigned to 36% of the virion proteins, which include structural proteins as well as enzymes for transcription, DNA repair, redox reactions and protein modification. Homologs of 36 CroV virion proteins have previously been found in the virion of *Acanthamoeba polyphaga* mimivirus. The overlapping virion proteome of CroV and Mimivirus reveals a set of conserved virion protein functions that were presumably present in the last common ancestor of the *Mimiviridae*.

Introduction

Cafeteria roenbergensis virus (CroV) is a double-stranded (ds) DNA virus that infects the marine unicellular heterotrophic nanoflagellate *C. roenbergensis* (Fenchel and Patterson, 1988), and is the first giant virus reported to infect zooplankton. It was isolated from a sample of concentrated viruses collected from several locations in the western Gulf of Mexico, although the host was originally identified as *Bodo* sp. (Garza and Suttle, 1995). The genome of CroV is approximately 730 kilobase pairs (kbp), 617 kbp of which has been assembled into a linear chromosome with a nucleotide composition of 76.7% A+T (Fischer et al., 2010). The remaining ~110 kbp is presumed to consist of highly repetitive DNA, with hundreds of copies of the ~66 bp long FNIP/IP22 repeat motif (O'Day et al., 2006), which are found in several regions of the CroV genome and increase in density towards the ends of the CroV assembly (Fischer et al., 2010). The missing genome parts are thus unlikely to contain unique coding information. The assembled part of the genome is predicted to contain at least 544 protein-coding sequences (CDS), including DNA replication and transcription proteins as well as enzymes involved in DNA repair, glycosylation, ubiquitination, and other processes. CroV is the sole member of the *Cafeteriavirus* genus in the *Mimiviridae* family within the proposed order *Megavirales* (Colson et al., 2013), which is synonymous to the Nucleo-Cytoplasmic Large DNA Virus (NCLDV) clade that comprises the families *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*, *Phycodnaviridae* and *Poxviridae* (Iyer et al., 2001; Yutin and Koonin, 2012). Like *Acanthamoeba polyphaga* mimivirus (Raoult et al., 2004), CroV reproduces in large cytoplasmic factories, where DNA replication, transcription, and particle assembly are thought to take place. In addition, the CroV virion factory is targeted by the Mavirus virophage, a small dsDNA virus with a 19 kbp genome that depends on CroV for its propagation (Fischer and Suttle, 2011).

The mature CroV virion consists of a 300 nm diameter outer protein shell with icosahedral symmetry, an underlying lipid membrane, and an inner core that contains the genome (Fig. S1, C. Xiao, M.G. Fischer, C.A. Suttle, unpublished results). The virion is the physical entity of a virus; it protects the genetic material during the extracellular stage of the viral life cycle and delivers it to a suitable host cell. Knowledge of the protein composition of a virion can thus yield important clues about the extracellular, penetration, and early intracellular stages of the virus replication cycle. Mass spectrometry-based techniques have been applied to elucidate the virion protein sets of several large and giant DNA

[*] Correspondence to: Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany. Tel.: +49 6221 486506.
E-mail address: mfischer@mpimf-heidelberg.mpg.de (M.G. Fischer).
[1] Present address: Quebec Genomics Center, 2705 boul. Laurier, R2-2710 Sainte-Foy (Québec), Canada G1V 4G2.

http://dx.doi.org/10.1016/j.virol.2014.05.029
0042-6822/© 2014 Elsevier Inc. All rights reserved.

viruses (Allen et al., 2008; Renesto et al., 2006; Dunigan et al., 2012; Resch et al., 2007; Wang et al., 2010; Song et al., 2006). For instance, 28 viral proteins were identified in the virion of the coccolithovirus EhV-86 (Allen et al., 2008), the baculovirus AcMNPV virion was found to consist of 34 viral proteins (Wang et al., 2010), 44 viral proteins were reported for Singapore grouper iridovirus (Song et al., 2006), and poxvirus virions contain about 80 viral proteins (Resch et al., 2007; Chung et al., 2006; Yoder et al., 2006). The giant Mimivirus was found to package 114 viral proteins (Renesto et al., 2006) although a subsequent re-analysis (Claverie et al., 2009) added another 23 statistically well supported proteins for a total of at least 137. With 148 viral proteins, the phycodnavirus PBCV-1 also packages a large number of proteins (Dunigan et al., 2012). Interestingly, the virion proteomes of large and giant dsDNA viruses often comprise a significant number of proteins that are predicted not to contribute to the virion architecture, but to catalyze various enzymatic reactions, e.g. mRNA synthesis (Renesto et al., 2006; Dunigan et al., 2012; Resch et al., 2007). In this study, we analyzed the protein composition of the CroV virion using an LC–MS/MS proteomic approach and compared the CroV virion proteome to that of the related Mimivirus.

Results and discussion

The proteomic analysis of CroV strain BV-PW1 reveals a virus with a complex repertoire of proteins that is predicted to comprise not only structural proteins, but also a broad suite of proteins typically associated with cellular processes. These include proteins predicted to be involved in mRNA synthesis, DNA repair, disulfide bond formation, as well as various proteases and phosphatases, and the first report of a mechanosensitive ion channel protein encoded by a virus. These results and their implications are discussed in detail below.

Identification of proteins in the CroV particle by LC–MS/MS

Density gradient ultracentrifugation was used to purify CroV particles for protein analysis. The resulting virion preparation was free of cells and other visible contaminants, as determined by epifluorescence microscopy. The proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and digested with trypsin in-gel. Analysis of the peptides by LC–MS/MS and subsequent comparison of the fragment spectra with predicted CDSs and unidentified reading frames (URFs) from CroV resulted in the unambiguous identification of 141 CroV proteins (Table 1, SI file 1). We also identified two proteins from the Mavirus virophage, the major capsid protein MV18 and the predicted hydrolase MV13 (Fig. 1, SI file 1). A total of 60 host cell proteins were identified by querying the transcriptome peptide data of *C. roenbergensis* strain E4-10, available through the Camera metagenomics portal at http://camera.calit2.net/mmetsp/details.php?id=MMETSP0942 as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling et al., 2014) (Fig. 1, SI files 1 and 2). Additional host proteins may have remained unidentified because the *C. roenbergensis* genome has not been sequenced. In order to estimate the relative protein quantities in the analyzed sample, intensity values (IntVals) were calculated from the average intensity of the five-most intense ions for each protein (Table 1, SI file 1). The distribution of IntVals is shown in Fig. 1. For better direct comparison, the IntVals in Table 1 are also expressed as percentages of the IntVal of the major capsid protein, which was the most abundant protein in the dataset.

The predicted molecular weight of the 141 CroV virion proteins ranged from 7 to 217 kDa and the isoelectric points ranged from 3.7 to 11.1. 50% of the virion proteins had a molecular weight (MW) of less than 30.5 kDa or a pI greater than 8.4. This reflects the overall predicted protein spectrum encoded by CroV, with 50% of CroV proteins having a MW below 28.5 kDa and a pI greater than 8.8. Hence, the slight trend towards small basic proteins does not result from preferential packaging of these proteins (Fig. S2). Some of the small basic proteins may be involved in DNA binding, as proposed for PBCV-1 (Dunigan et al., 2012). The strand distribution of virion protein-coding genes was slightly biased (61%) towards one strand (forward strand of the reference genome NC_014637). Ninety (64%) of the 141 CroV-encoded virion proteins had no functional annotation; hence their role in the virion and during the viral infection cycle is unknown. The remainder of the virion proteins could be classified into the following functional categories: structure, transcription, DNA repair, redox control, protein modification, and miscellaneous (Table 1). Gene Ontology (GO) enrichment analysis confirmed that proteins involved in virion structure, transcription, DNA repair, and redox control were significantly overrepresented in the virion proteome compared to the full CroV proteome (SI file 3, Fig. 2). On the other hand, none of the 10 predicted Ubiquitin pathway-associated proteins or the 29 FNIP/IP22 repeat-containing proteins were found in the virion.

Structural virion proteins

Proteins found in a mature virus particle are typically classified as “structural”. Mature particles of giant DNA viruses, however, contain proteins that serve non-structural roles (e.g. transcription enzymes). Therefore, we use the term “structural” to refer to proteins that are predicted to contribute to the structural integrity of the virion. This applies to capsid- and core-associated proteins, as well as some transmembrane proteins interacting with the internal lipid layer of the virion. Non-structural virion proteins, on the other hand, catalyze enzymatic reactions that occur early during the infection cycle or possibly even within the virion itself, but are not essential for the virion architecture. All four predicted CroV capsid proteins were identified in this study. The major capsid protein (MCP) crov342 had the highest IntVal of all detected proteins (1.16E+11), followed by the core protein crov332 (IntVal=7.82E+10, Fig. 1, Table 1). Capsid proteins 2-4 (crov398, crov321, crov176) had low IntVals ranging from 5.03E+06 to 4.55E+08 and are therefore minor virion constituents. If we assume that one CroV virion contains roughly 15,000 copies of the MCP (Chuan Xiao, University of Texas El Paso, personal communication), there would be ~10,000 copies of the core protein, ~60 copies of capsid protein 2, and 1 and 2 copies of capsid proteins 3 and 4, respectively, per virion. Another predicted structural component of the virion is the phage tail collar domain-containing protein crov148 (IntVal=4.15E+09). Given the size and complexity of the mature CroV particle, it is likely that additional proteins are required for its structural integrity. We therefore considered proteins of unknown function as candidates for structural proteins if they had more than 50% sequence coverage and IntVals higher than that of the most abundant, clearly non-structural protein in our dataset (crov224, DNA-directed RNA polymerase Rpb2, IntVal=2.76E+09, see Fig. 1, SI file 1). The identification criteria were chosen conservatively, in order to minimize the inclusion of potential non-structural proteins, such as yet unidentified transcription components. Seven CDSs of unknown function matched these criteria: crov039, crov174, crov175, crov187, crov318, crov320 and crov330 (Table 1, Fig. 1). Homologs of crov187, crov318 and crov330 are present in other *Mimiviridae* members and, in the case of Mimivirus, were identified as virion proteins. Coding sequence crov330 appears to be a CroV-specific gene fusion, as its orthologs in all other *Mimiviridae* members are encoded by two adjacent CDSs. The remaining four CDSs are unique to CroV. Two of the seven potentially new structural proteins (crov240 and crov320) contain transmembrane domains (TMDs) and may interact with the internal lipid membrane of the CroV virion. A total of 39 proteins (28%) identified in this study are predicted to contain one or more TMDs (Table 1). This is far less than the 82% (23 of 28) *Emiliania huxleyi* virus-86 virion proteins that are predicted to contain TMDs (Allen et al., 2008). However, EhV-86 contains an outer membrane which it acquires during budding from the host membrane (Mackinder et al., 2009), and may therefore need to anchor proteins in its lipid envelope to mediate host recognition and cell entry. In contrast, only 12% of virion proteins in Mimivirus, which does not contain an outer lipid envelope, are predicted TMD proteins based on TMHMM analysis of the Mimivirus virion proteome (Renesto et al., 2006). Overall, CroV encodes 70 predicted TMD proteins, 39 of which are present in the virion. Proteins with predicted TMDs are therefore overrepresented in the CroV virion (Fisher's exact test, *p*-value 1.5E-08).

Non-structural virion proteins

Transcription-related proteins constituted the largest class of predicted non-structural proteins that were associated with the CroV virion. The packaged transcription machinery consisted of eight DNA-directed RNA polymerase II subunits (Rpb1, 2, 3/11, 5, 6, 7, 9, 10), two predicted early transcription factors (vaccinia virus A7- and D6-like), two vaccinia virus D11-like transcription factors, the trifunctional mRNA capping enzyme, polyadenylate polymerase catalytic subunit, DNA topoisomerase IB, and an RNA helicase (Table 1). The Rpb2 subunit crov224 contains a 146 amino acid long intein (self-splicing protein intron) insertion in its RNA polymerase beta chain signature motif (http://tools.neb.com/inbase/intein.php?name=CroV+RPB2). The Rpb2 intein exhibits all the conserved residues necessary for self-splicing; however, there are no data demonstrating that the intein is functional and able to remove itself from the host protein (extein) after translation. In this study, we found 54 peptides spanning 46% of the Rpb2 precursor sequence. None of these peptides contained intein residues and, the intein sequence represented the longest contiguous region for which no peptide match was found (Fig. 3); this may indicate the absence of the intein from the virion-associated Rpb2 protein. On the other hand, no peptide crossing the extein–extein junction was found either, which would have proven a successful intein splicing event. However, the theoretical trypic peptide spanning the splice junction (amino acid sequence FSTK) would have been too short to be unambiguously identified (Fig. 3). Overall, the virion presence of a large set of transcription proteins strongly suggests that CroV is able to initiate gene transcription within the viral core immediately after host cell entry, in a manner similar to vaccinia virus and Mimivirus (Broyles, 2003; Mutsafi et al., 2010).

Six of the identified virion proteins were predicted DNA repair enzymes. The putative DNA photolyase, crov149, may be involved in intra-virion DNA photorepair, as has been demonstrated for Fowlpox virus (Srinivasan and Tripathy, 2005). Remarkably, the CroV virion also contains the following enzymes predicted to function in the base excision repair (BER) pathway: a Nei-like DNA glycosylase (crov303) to excise a damaged DNA base, an apurinic/apyrimidinic (AP) endonuclease (crov106) to cleave the sugar-phosphate backbone at the resulting AP site, a family X DNA polymerase (crov458) to fill in the missing nucleotide and a NAD-dependent DNA ligase (crov462) to seal the nicked DNA strand. The finding that BER enzymes are packaged in the CroV particle implies that they might be involved in pre-replicative DNA repair (Redrejo-Rodríguez and Salas, 2014).

The CroV virion contained six predicted redox-active proteins: three thiol oxidoreductases, two protein disulfide isomerases (PDIs), and a glutaredoxin. Enzymes of the PDI family catalyze the formation of disulfide bonds in other proteins and are involved in a variety of cellular processes, such as oxidative protein folding and antigen presentation (Appenzeller-Herzog and Ellgaard, 2008). Disulfide bonds also play an important role in many viruses, e.g. during virus entry and for stabilization of structural virion proteins. For instance, simian virus 40 uncoating is mediated by a cellular oxidoreductase (Schelhaas et al., 2007). In HIV-1, a PDI is probably required for virion entry, as it cleaves two disulfide bonds of the envelope glycoprotein gp120 which interacts with the CD4 receptor, inducing virus–cell fusion (Ryser et al., 1994). Although CroV does not have an external lipid envelope, PDIs could nevertheless play a role in CroV–cell membrane fusion events, e.g. to release the CroV core from a hypothetical endocytic vesicle after virion entry. Another possibility is that these redox proteins could assist in the folding of structural proteins within the cytoplasmic virion factory.

The functional group of protein modification was represented by four peptidases, two protein phosphatases, and a protein kinase. Among the remaining CroV-encoded virion proteins with a functional annotation is crov217, which contains a glycosyltransferase domain near its amino terminus and lacks homologs in other viruses. It was the fourth-most abundant protein found in this study, with an IntVal of 9.75E+09, 46 unique peptides, and 61% coverage, meaning the virion contained roughly one copy of the crov217 protein per 12 copies of the MCP. Despite its annotation as a glycosyltransferase, crov217 may thus serve a structural role in the virion. Also noteworthy is crov294, the first viral homolog of a mechanosensitive ion channel protein. Mechanosensitive channels of large conductance (MscL) are most commonly found in bacteria, but also occur in some archaea and eukaryotes. These channel proteins are involved in osmoregulation and open in response to mechanical stimuli, e.g. membrane stretching induced by osmotic shock (Booth and Blount, 2012). The $\mathrm{MscL}_{\mathrm{CroV}}$ protein is 101 amino acids long and has two predicted TMDs (Table 1). Based on its IntVal of 4.51E+08, the virion contains approximately 60 copies of this protein, which corresponds to 12 pentamers, the native quarternary configuration of MscL proteins. $\mathrm{MscL}_{\mathrm{CroV}}$ is predicted to be located in the internal virion membrane where it may play a role in osmoregulation of the CroV core.

Correlation of virion proteins with CroV gene promoter elements

Genes encoding virion structure components tend to be expressed late during infection (Resch et al., 2007). In the immediate 5′ upstream region of CroV CDSs there are sequence motifs that are predicted to govern early and late gene expression, based on the correlation of these sequence motifs with gene expression data obtained from a custom CroV-specific microarray (Fischer et al., 2010). Considering that no expression was found for 56 CDSs whose proteins were detected in the virion, the microarray study clearly underestimated the total number of viral genes expressed during CroV infection. The CroV early gene promoter motif contains the conserved sequence AAAAATTGA, for which the first adenine residue is typically located about 42 nucleotides upstream of the predicted start codon (position -42) (Fig. 4A and C). The early motif is shared with Mimivirus, where it was originally identified (Suhre et al., 2005). The CroV late gene promoter contains a TCT[A/T] core motif flanked by AT-rich sequences, and its first thymine residue is typically located around position -14 (Fig. 4B and C).

An updated analysis of the promoter motifs (Fig. 4, Table S1) showed that out of the 544 CroV CDSs, 182 (33%) were preceded by the early promoter motif, 96 (18%) were preceded by the late promoter motif, 13 (2%) were preceded by both motifs, and 253 (47%) had no recognizable promoter motif in their 5′ upstream

Table 1
CroV virion proteins identified by LC–MS/MS. The virion proteins are grouped by predicted function: Blue, structure; green, transcription; red, DNA repair; yellow, redox control; purple, protein modification; gray, miscellaneous; white, function unknown.

CDS[a]	Putative fuction	MW[b]	Intrnsity	%Int[c]	Pep[d]	Cov[e]	TM[f]	pI	Prom[g]	Expr[h]	Mimivirus[i]
342	Major capsid protein	57.24	1.16E+11	100	29	70.2	0	5.4	Late	6	**L425**
398	Capsid protein 2	53.58	4.55E+08	0.39	16	40.6	0	6.9	Late	6	R441
321	Capsid protein 3	76.63	5.03E+06	< 0.01	1	2.3	0	5.3	Late	N/D	R439
176	Capsid protein 4	58.77	1.62E+07	0.01	6	12.8	0	7.9	–	N/D	R440
332	Core protein	67.33	7.82E+10	67.24	43	80.5	0	8.7	Late	6	**L410**
148	Phage tail collar domain protein	26.22	4.15E+09	3.57	10	53.3	0	6.2	Late	6	–
39	Unknown	17.75	9.42E+09	8.10	8	66.7	0	9.9	Late	N/D	–
174	Unknown	46.85	5.11E+09	4.40	28	59.1	0	7.0	Late	6	–
175	Unknown	28.07	7.62E+09	6.55	14	64.2	1	9.6	Late	N/D	–
187	Unknown	14.42	3.97E+09	3.42	7	63.5	0	8.8	Late	12	**R459**
318	Unknown	9.33	4.39E+09	3.78	8	91.0	0	7.3	Late	N/A	**L485**
320	Unknown	20.00	1.37E+10	11.79	8	52.7	3	5.1	Late	6	–
330	Unknown	125.93	4.50E+09	3.87	46	50.6	0	9.6	–	6	**R402/R403**
368	DNA-directed RNA polymerase II, subunit Rpb1	169.46	2.26E+09	1.95	52	36.8	0	7.3	Early	2	**R501**
224	DNA directed RNA polymerase II, subunit Rpb2 (intein-containing)	152.40	2.76E+09	2.37	54	46.4	0	8.5	Early	2	**L244**
482	DNA-directed RNA polymerase II, subunit Rpb3/Rpb11	37.93	6.73E+08	0.58	8	26.8	0	5.2	Early	0	**R470**
439	DNA-directed RNA polymerase II, subunit Rpb5	22.36	2.75E+08	0.24	8	34.7	0	10.2	Early	6	**L235**
491	DNA-directed RNA polymerase II, subunit Rpb6	17.71	8.56E+07	0.07	4	19.6	0	4.2	Early	2	R209
437	DNA-directed RNA polymerase II, subunit Rpb7	27.15	1.42E+08	0.12	5	31.5	0	5.2	Early, late	2	**L376**
492	DNA-directed RNA polymerase II, subunit Rpb9	21.27	3.07E+08	0.26	6	27.6	0	9.4	Early	1	**L208**
201	DNA-directed RNA polymerase II, subunit Rpb10	9.07	7.49E+07	0.06	3	46.8	0	7.8	Early, late	N/A	R357b
109	VV D6-like very early transcription factor, small subunit	132.21	7.82E+08	0.67	28	28.1	0	9.2	Late	6	**L377**
292	VV A7-like very early transcription factor, large subunit	217.13	3.44E+08	0.30	22	15.0	0	9.5	–	1	**R326/R327**
131	VV D6/D11-like transcription factor	66.25	1.95E+07	0.02	3	6.7	0	10.0	–	N/C	**R563**
283	VV D11-like transcription termination factor	90.87	2.37E+09	2.04	39	52.3	0	9.3	Late	3	**R350**
212	mRNA capping enzyme	119.40	9.97E+08	0.86	29	30.3	0	8.9	–	3	**R382**
286	Poly(A) polymerase catalytic subunit	53.72	5.37E+08	0.46	17	45.3	0	9.0	–	host	**R341**
152	DNA topoisomerase IB	62.14	7.63E+07	0.07	8	17.4	0	10.3	–	N/C	**R194**
118	RNA helicase	193.75	3.51E+07	0.03	6	3.7	0	9.4	–	N/D	**R366**
149	DNA photolyase	58.23	7.23E+06	0.01	2	5.6	0	10.1	–	N/D	R852/R853/R855
303	Formamidopyrimidine-DNA glycosylase	33.30	1.98E+07	0.02	3	9.7	0	9.9	–	6	L315
106	Apurinic-apyrimidinic endonuclease 1	30.55	3.45E+06	< 0.01	1	5.0	0	9.7	Early	2	–
458	Family X DNA polymerase	40.98	3.53E+07	0.03	6	18.6	0	10.2	Early	3	**L318**
462	NAD-dependent DNA ligase	74.84	2.65E+07	0.02	4	6.5	0	10.2	Early	2	R303
200	3′-5′ exonuclease	41.72	1.89E+08	0.16	9	29.5	0	9.8	–	N/D	**L533**
13	N-terminal oxidoreductase domain	55.96	6.93E+07	0.06	8	17.7	1	9.8	–	N/D	–
143	Erv1/Alr family thiol oxidoreductase	21.36	2.88E+08	0.25	4	23.2	0	9.3	–	6	**R596**
179	Glutaredoxin	13.18	6.79E+07	0.06	4	38.1	0	10.3	Late	N/D	–
444	Erv1/Alr family thiol oxidoreductase	19.29	2.01E+08	0.17	5	43.6	1	9.5	–	N/A	R368
172	Protein disulfide isomerase	17.49	9.94E+08	0.86	9	52.9	1	6.5	Late	N/D	**R362**
379	Protein disulfide isomerase/thioredoxin	13.93	8.12E+07	0.07	4	27.1	0	9.9	Late	N/D	**R443**
67	Metallodopeptidase (M13 family)	72.68	6.66E+07	0.06	7	12.2	0	6.1	–	6	R519
137	Insulinase-like metalloprotease	102.12	7.14E+07	0.06	9	11.3	0	9.5	–	0	L233
184	Cysteine protease	34.71	8.56E+07	0.07	6	26.5	0	8.7	–	6	**R355**
309	Serine/threonine protein kinase	46.62	6.33E+07	0.05	5	15.1	0	9.7	–	0	**R400**
475	Family 2C serine/threonine phosphatase	30.50	2.20E+08	0.19	7	34.9	0	8.7	Late	6	R306/**R307**
485	OTU family cysteine protease	31.14	2.39E+08	0.21	4	19.1	0	5.9	–	N/D	–
525	Serine/threonine protein phosphatase	73.18	8.95E+07	0.08	11	19.8	0	6.5	Late	0	–
14	DoxX family membrane protein	14.28	3.93E+07	0.03	2	19.8	3	10.0	–	N/D	–
136	Putative kinase	63.94	3.66E+08	0.31	12	24.6	0	9.4	Late	6	–
171	ADP-ribosyl glycohydrolase	39.91	6.02E+07	0.05	2	6.7	0	6.5	Early	N/D	L444
217	Branch superfamily glycosyl transferase	101.45	9.75E+09	8.39	46	60.8	0	9.4	Late	6	–
232	ABC-type amino acid transporter	56.94	1.66E+08	0.14	5	12.8	3	8.1	Early	0	–
294	MscL superfamily membrane protein (ion channel)	11.68	4.51E+08	0.39	4	41.6	2	5.8	Late	N/D	–

Table 1 (*continued*)

CDS[a]	Putative fuction	MW[b]	Intrnsity	%Int[c]	Pep[d]	Cov[e]	TM[f]	pI	Prom[g]	Expr[h]	Mimivirus[i]
302	AAA+ family ATPase	58.23	1.92E+06	< 0.01	1	3.8	1	9.5	Early	0	L573
313	MPP-superfamily metallophosphatase	36.92	1.46E+07	0.01	2	7.3	0	8.1	Late	2	**R398**
319	Ankyrin repeat-containing protein	165.57	3.04E+09	2.62	41	33.6	0	5.6	Late	N/D	**L484**
417	SET domain-containing protein	18.12	8.26E+07	0.07	2	16.1	0	7.7	–	N/D	L678
442	FtsJ-like methyltransferase	52.42	4.39E+08	0.38	12	29.7	0	9.5	Early	24	**R383**
456	Ankyrin repeat-containing protein	72.26	3.41E+07	0.03	6	9.5	0	8.3	–	N/D	L371
484	Ribonuclease H1	19.00	1.13E+07	0.01	1	6.7	0	10.2	Late	N/D	R298/R299
19	Unknown	58.44	4.76E+08	0.41	21	43.1	0	7.8	–	N/D	–
26	Unknown	34.99	1.88E+07	0.02	4	16.0	0	9.7	–	N/D	–
34	Unknown	19.37	4.01E+07	0.03	4	26.3	0	11.0	–	N/D	–
47	unknown	106.45	6.07E+07	0.05	4	6.4	0	4.2	Late	6	–
60	Unknown	28.49	1.57E+08	0.14	3	15.4	1	7.7	Late	48	–
81	Unknown	8.05	6.62E+06	0.01	1	11.1	2	6.5	Late	N/A	–
83	Unknown	28.05	2.64E+07	0.02	1	3.3	8	6.9	–	N/D	–
85	Unknown	11.20	1.01E+07	0.01	2	32.3	1	8.0	Late	N/A	–
88	Unknown	80.46	4.20E+07	0.04	7	11.4	1	4.8	Late	72	–
89	Unknown	49.32	3.40E+06	< 0.01	2	5.1	1	4.7	–	6	–
90	Unknown	45.66	1.79E+08	0.15	9	25.9	1	4.2	Late	N/D	–
094	Unknown	21.80	3.03E+09	2.61	7	29.7	0	11.1	Late	6	–
098	Unknown	9.04	1.37E+07	0.01	2	21.5	0	10.4	Late	N/A	–
110	Unknown	36.21	1.40E+09	1.20	13	46.5	0	6.7	–	2	–
132	Unknown	58.07	2.87E+08	0.25	9	20.6	0	9.5	Late	0	–
133	Unknown	55.72	3.13E+08	0.27	12	27.4	0	8.3	Late	N/C	–
141	Unknown	18.78	4.89E+06	< 0.01	1	16.2	1	4.9	–	N/D	–
142	unknown	210.28	4.21E+08	0.36	9	6.4	0	4.4	–	6	–
145	Unknown	29.24	1.13E+09	0.97	9	51.8	0	10.1	Late	N/D	–
146	Unknown	30.45	9.54E+06	0.01	1	4.7	0	8.9	Late	N/D	–
150	Unknown	94.17	4.49E+06	< 0.01	3	4.7	0	9.0	–	3	–
165	Unknown	21.79	3.99E+07	0.03	4	26.6	0	4.9	–	N/D	–
167	Unknown	14.09	5.89E+08	0.51	3	36.1	1	8.9	Late	N/D	–
173	Unknown	32.14	7.29E+08	0.63	11	44.9	0	9.5	Late	N/D	–
185	Unknown	98.74	4.33E+08	0.37	17	25.4	0	10.0	Late	6	**L454**
186	Unknown	22.12	1.54E+09	1.32	18	87.6	0	7.5	Late	N/D	**R457**
192	Unknown	20.20	8.85E+06	0.01	1	8.8	0	9.8	–	N/D	–
194	Unknown	22.14	4.36E+07	0.04	3	14.3	2	7.5	Late	12	R468
199	Unknown	81.58	1.07E+09	0.92	18	27.1	0	7.5	Late	6	–
202	Unknown	12.16	5.32E+08	0.46	5	40.8	1	5.2	–	N/D	–
213	Unknown	16.81	3.37E+07	0.03	2	17.3	0	10.7	Late	N/D	–
215	Unknown	20.73	7.39E+08	0.64	3	9.4	0	6.7	–	6	**L492**
222	Unknown	45.85	3.79E+08	0.33	11	32.7	1	8.2	–	0	–
225	Unknown	35.01	3.70E+07	0.03	3	13.4	0	8.9	–	N/D	–
227	Unknown	20.27	4.29E+06	< 0.01	1	4.4	1	3.7	–	N/D	–
228	Unknown	23.76	1.85E+09	1.60	5	25.2	1	3.7	Late	6	**R489**
240	Unknown	8.37	1.62E+09	1.39	2	23.1	1	5.6	Late	N/A	–
288	Unknown	21.82	2.10E+08	0.18	5	26.9	0	8.4	Late	6	–
289	Unknown	7.19	6.04E+06	0.01	1	24.6	0	8.5	–	N/A	–
290	Unknown	35.37	2.87E+06	< 0.01	1	4.6	0	7.2	Late	6	–
293	Unknown	8.69	9.58E+05	< 0.01	1	20.3	0	6.8	Late	N/A	–
295	Unknown	20.33	2.33E+07	0.02	3	20.5	0	6.5	–	3	R329
296	Unknown	23.19	1.96E+08	0.17	8	49.5	0	4.7	Late	N/D	–
310	Unknown	34.43	2.33E+08	0.20	4	19.5	1	7.7	–	N/D	–
311	Unknown	183.08	2.20E+09	1.89	48	29.7	0	8.6	–	N/D	**R472**
312	Unknown	39.44	7.62E+08	0.66	4	8.9	0	11.0	Late	6	–
324	Unknown	14.21	1.54E+07	0.01	1	9.3	0	9.5	Late	N/D	–
326	Unknown	75.90	2.37E+07	0.02	2	3.5	0	5.3	–	N/D	R481
327	Unknown	38.75	3.58E+07	0.03	4	18.1	1	4.1	Late	6	–
331	Unknown	20.53	5.22E+07	0.04	4	24.6	0	9.6	–	N/D	R409
334	Unknown	17.04	1.49E+07	0.01	1	5.6	0	8.7	–	N/D	–
335	Unknown	11.72	1.45E+07	0.01	1	11.4	1	5.1	Late	N/D	–
336	Unknown	51.50	1.92E+09	1.65	13	32.1	0	7.9	Late	6	**L417**

340	Unknown	11.33	9.22E+07	0.08	4	40.6	1	4.6	–	N/A	–
358	Unknown	39.45	1.62E+08	0.14	3	13.6	0	9.9	Late	0	–
361	Unknown	20.69	6.95E+07	0.06	3	33.0	2	7.4	Late	6	R335
363	Unknown	39.99	4.10E+08	0.35	8	23.7	1	4.7	Late	host	–
365	Unknown	14.38	4.76E+08	0.41	2	18.4	3	6.7	Late	6	–
366	Unknown	8.56	2.83E+07	0.02	2	23.9	1	8.7	Late	N/A	–
390	Unknown	119.10	1.08E+07	0.01	2	1.9	0	3.8	–	6	–
399	Unknown	43.39	4.16E+05	< 0.01	1	2.7	0	4.4	–	6	–
400	Unknown	41.43	1.44E+09	1.24	16	61.0	0	6.8	Late	N/D	–
401	Unknown	10.57	3.24E+07	0.03	3	37.5	1	9.6	–	N/A	–
404	Unknown	15.54	8.88E+08	0.76	4	34.5	2	8.4	Late	6	–
410	Unknown	12.71	1.47E+08	0.13	5	55.0	0	10.0	–	0	–
411	Unknown	17.90	3.27E+08	0.28	5	33.8	1	8.8	Late	12	–
419	Unknown	56.65	2.91E+08	0.25	14	35.6	0	9.2	Late	N/D	–
425	unknown DUF2177	15.13	2.29E+08	0.20	2	14.4	4	5.8	–	N/D	–
432	Unknown	30.99	1.54E+08	0.13	6	28.7	0	10.3	Late	3	–
433	Unknown	15.41	6.47E+08	0.56	5	41.4	0	6.9	–	0	–
434	Unknown	6.70	5.10E+07	0.04	2	37.0	0	10.2	–	N/A	–
436	Unknown	16.10	1.07E+09	0.92	7	37.6	0	9.5	Late	N/D	–
451	Unknown	24.94	2.31E+08	0.20	3	15.2	1	9.1	Late	6	–
453	Unknown DUF4419	40.93	3.19E+07	0.03	4	15.0	0	6.4	–	6	L662
465	Unknown	13.09	4.06E+07	0.03	2	21.4	0	5.1	Late	N/D	–
466	Unknown	23.42	1.04E+08	0.09	3	25.5	1	5.4	Late	N/D	–
487	Unknown	12.04	3.07E+08	0.26	7	75.8	0	10.2	–	N/D	–
488	Unknown	16.41	5.68E+07	0.05	3	33.3	0	10.1	Late	N/D	–
496	Unknown	25.14	9.20E+07	0.08	3	15.3	1	4.5	Late	N/D	–
498	Unknown	173.25	1.59E+09	1.37	35	25.9	0	7.6	–	0	–

[a] aCroV gene number.

[b] Predicted molecular weight in kilodalton.

[c] Relative intensity values, expressed as percentages of major capsid protein intensity.

[d] Number of unique peptides found per protein.

[e] Percentage of the protein covered by identified peptides.

[f] Number of transmembrane helices predicted by TMHMM v2.0.

[g] Indicates whether the CDS has one of the two identified CroV gene promoter sequences in its 5′ upstream region (early: AAAAATTGA, late: TCT[T/A]).

[h] Onset of gene expression in hours post infection as determined in a previous microarray experiment (N/A: data not available; N/C:expression status unclear; N/D: transcript not detected; host: microarray probes for this transcript cross-hybridized with a host cell transcript.

[i] Homologous CDSs in Acanthamoeba polyphagamimivirus with virion proteins in bold and underlined. In case of several homologs in Mimivirus, only the best hit is shown. Multiple listings for Mimivirus represent cases where the protein is encoded by a single CDS in CroV, but two or three adjacent CDSs in Mimivirus.

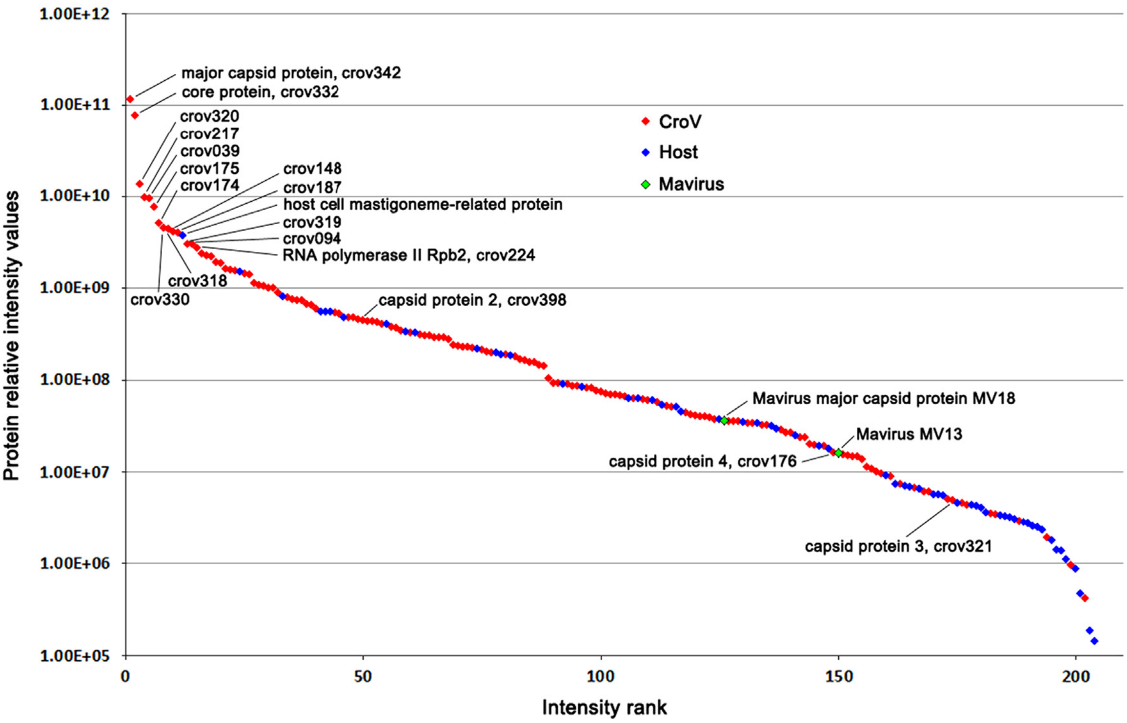


Fig. 1. Rank vs. intensity values chart of the 141 CroV virion proteins, 60 host cell proteins, and 2 Mavirus proteins identified in this study. Select proteins of interest are labeled.

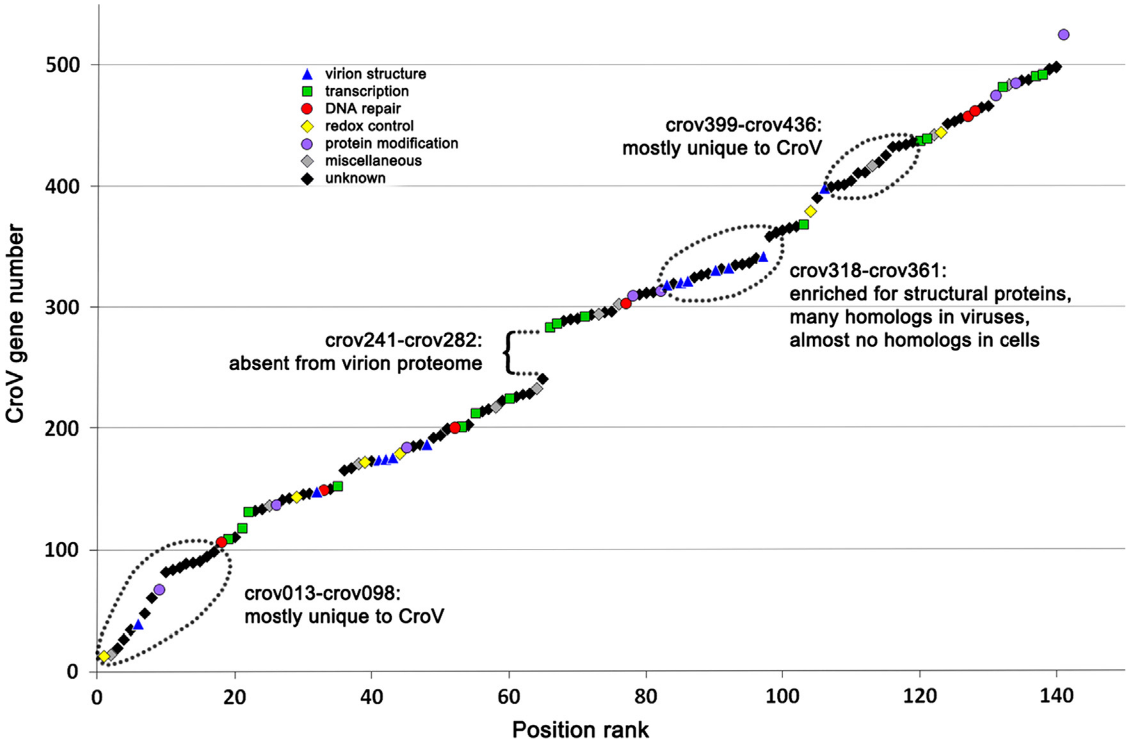


Fig. 2. Rank vs. relative genome position chart of the 141 CroV virion proteins. The virion proteins are plotted in increasing order of their gene number, which reflects their relative location in the CroV genome from 5′ end (left) to 3′ end (right). The predicted functional protein categories are indicated by different colors.

region. We observed a significant bias among genes encoding virion proteins with respect to the distribution of early and late promoter motifs (Fig. 4D). Whereas only 15 of the 195 (8%) CDSs associated with the early promoter motif encoded a virion protein (under-representation of CDSs with early promoter, *p*-value 2.2E-14, Fisher's exact test), this was the case for 72 of the 109 (66%) late promoter-associated CDSs (over-representation of CDSs with late promoter, *p*-value 7.0E-24, Fisher's exact test). Therefore, proteins originating from genes preceded by the late promoter motif are preferentially packaged in the CroV virion. Interestingly, 11 of the 15 virion proteins linked to the early promoter motif are annotated as transcription or DNA repair proteins; the remaining four "early" proteins also have non-structural annotations (FtsJ-like methyltransferase, ADP-ribosyl glycohydrolase, ABC-type amino acid transporter, and AAA+ family ATPase). None of the 82 virion proteins of unknown function were associated with the early promoter motif; whereas, 47 (57%) virion proteins of unknown function were encoded by a gene preceded by the late promoter motif.

Exclusion of a 38 kbp genomic region from the proteome

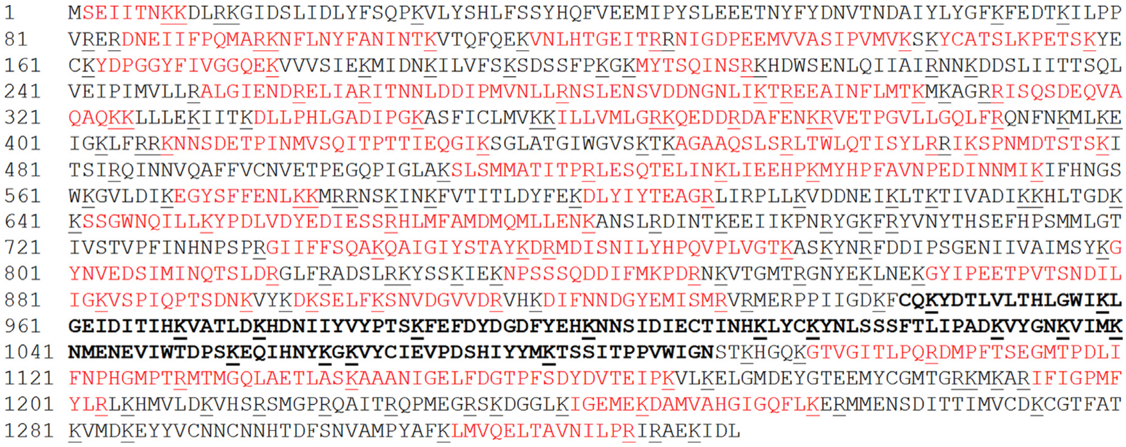
Overall, the coding sequences for proteins found in the proteome of the CroV virion were distributed rather uniformly across the CroV genome (Fig. 2); however, there were some interesting patterns. Most of the virion proteins from crov013 to crov098 and crov399 to crov436 lack homologs in other viruses, and thus may represent host-adapted proteins or structural components specific to the CroV virion architecture. In contrast, 10 of the 17 proteins from crov318 to crov361 had homologs in other *Mimiviridae* members, and included the major capsid protein and the core protein. These proteins thus may be responsible for structural features that are common to all *Mimiviridae* members.

None of the proteins from crov241 to crov282 were detected, as indicated by the gap in the plot in Fig. 2. This 38 kbp genomic region (between positions 265 kbp and 303 kbp) is devoid of the conserved early and late gene promoter signals, lacks homologs in other giant viruses and the CDSs were most similar to those in bacteria and plants that code for enzymes involved in carbohydrate metabolism. Three of these genes (crov265–crov267) are predicted to catalyze the biosynthesis of 3-deoxy-ᴅ-manno-octulosonate (KDO), an essential cell wall component in plants and bacteria. Based on the differences between the 38 kbp region and the rest of the CroV genome, we hypothesized that this region could have been laterally transferred from a bacterium (Fischer et al., 2010). The absence of proteins corresponding to CDSs crov241 to crov282 further highlights the unusual status of this region of the genome and may hint at an intracellular role for these enzymes, e.g. they might be involved in viral protein glycosylation. Alternatively, the DNA may not encode proteins, although mRNAs were detected for 13 of the 42 CDSs during a DNA microarray study (Fischer et al., 2010), suggesting that at least these 13 proteins are produced during CroV infection.

CroV and Mimivirus share a conserved set of virion proteins

Virion proteomes of several NCLDV members have been analyzed, including vaccinia virus (Resch et al., 2007; Chung et al., 2006; Yoder et al., 2006; Randall et al., 2004), Singapore grouper iridovirus (Song et al., 2006), Mimivirus (Renesto et al., 2006; Claverie et al., 2009), and the phycodnaviruses EhV-86 (Allen et al., 2008) and PBCV-1 (Dunigan et al., 2012). Of these viruses, only vaccinia virus, PBCV-1 and Mimivirus virions have a similarly complex protein composition to CroV. At least 7 of the 80 vaccinia virus proteins that are part of the mature virion have homologs in CroV, and, except for the A18 helicase, are also found in the CroV proteome. In addition to six proteins with significant sequence similarity (major core protein P4, RNA polymerase subunits 1 and 2 [J6 and A24], and transcription factors A7, D6, and D11), the redox proteins E10 and G4 are probably functional homologs of the CroV proteins crov143 and crov379.

We compared the 141 virion proteins of CroV with 137 virion proteins of Mimivirus that were identified with good statistical support using LC–MS/MS and MALDI-TOF-MS (Renesto et al., 2006; Claverie et al., 2009). Sixty-two additional Mimivirus virion proteins that were identified by Claverie et al. (2009) with lower statistical support (*E*-value >0.01) were not considered for this analysis. Using a BlastP *E*-value cutoff of 1E-04 and after masking low complexity regions, we found that 55 of the 141 CroV virion proteins had a Mimivirus homolog (Table 1). The SET domain-containing proteins crov417 and MIMI_L678, as well as the DUF4419-containing proteins crov453 and MIMI_L662 were included in the list of homologs after visual inspection of the respective pairwise sequence alignments, even though the *E*-values were above 1E-04. In addition, the DNA topoisomerase IB proteins, crov152 and MIMI_R194, were considered functional homologs despite a lack of amino acid sequence conservation. Of the 58 CroV virion proteins with Mimivirus homologs, 36 had a homolog that was packaged in the Mimivirus particle (underlined Mimivirus CDSs in Table 1). Fig. 5 shows a side-by-side comparison of the virion proteomes of CroV and Mimivirus. Some CroV CDSs had Mimivirus homologs that were encoded by two or three adjacent CDSs; hence, the number of shared virion proteins in Fig. 5 corresponds to the number of CroV proteins with homologs in Mimivirus. Only six (17%) of the 36 shared virion proteins could not be linked to a predicted function or recognizable domain, which is remarkable considering that 57% of the CroV virion proteins, and about half of the virion proteins in Mimivirus, was of unknown function. Most of the packaged proteins of unknown function are therefore lineage-specific. The 14 transcription-related proteins comprised the largest functional category of shared virion proteins (Fig. 5), which reflects the finding that early gene expression in *Mimiviridae* members is initiated within the viral core immediately after entering the host cytoplasm (Mutsafi et al., 2010). The shared structural proteins include the major capsid protein, the core protein, and three of the seven structural protein



```
1     MSEIITNKKDLRKGIDSLIDLYFSQPKVLYSHLFSSYHQFVEEMIPYSLEEETNYFYDNVTNDAIYLYGFKFEDTKILPP
81    VRERDNEIIFPQMARKNFLNYFANINTKVTQFQEKVNLHTGEITRRNIGDPEEMVVASIPVMVKSKYCATSLKPETSKYE
161   CKYDPGGYFIVGGQEKVVVSIEKMIDNKILVFSKSDSSFPKGKMYTSQINSRKHDWSENLQIIAIRNNKDDSLIITTSQL
241   VEIPIMVLLRALGIENDRELIARITNNLDDIPMVNLLRNSLENSVDDNGNLIKTREEAINFLMTKMKAGRRISQSDEQVA
321   QAQKKLLLEKIITKDLLPHLGADIPGKASFICLMVKKILLVMLGRKQEDDRDAFENKRVETPGVLLGQLFRQNFNKMLKE
401   IGKLFRRKNNSDETPINMVSQITPTTIEQGIKSGLATGIWGVSKTKAGAAQSLSRLTWLQTISYLRRIKSPNMDTSTSKI
481   TSIRQINNVQAFFVCNVETPEGQPIGLAKSLSMMATITPRLESQTELINKLIEEHPKMYHPFAVNPEDINNMIKIFHNGS
561   WKGVLDIKEGYSFFENLKKMRRNSKINKFVTITLDYFEKDLYIYTEAGRLIRPLLKVDDNEIKLTKTIVADIKKHLTGDK
641   KSSGWNQILLKYPDLVDYEDIESSRHLMFAMDMQMLLENKANSLRDINTKEEIIKPNRYGKFRYVNYTHSEFHPSMMLGT
721   IVSTVPFINHNPSPRGIIFFSQAKQAIGIYSTAYKDRMDISNILYHPQVPLVGTKASKYNRFDDIPSGENIIVAIMSYKG
801   YNVEDSIMINQTSLDRGLFRADSLRKYSSKIEKNPSSSQDDIFMKPDRNKVTGMTRGNYEKLNEKGYIPEETPVTSNDIL
881   IGKVSPIQPTSDNKVYKDKSELFKSNVDGVVDRVHKDIFNNDGYEMISMRVRMERPPIIGDKFCQKYDTLVLTHLGWIKL
961   GEIDITIHKVATLDKHDNIIYVYPTSKFEFDYDGDFYEHKNNSIDIECTINHKLYCKYNLSSSFTLIPADKVYGNKVIMK
1041  NMENEVIWTDPSKEQIHNYKGKVYCIEVPDSHIYYMKTSSITPPVWIGNSTKHGQKGTVGITLPQRDMPFTSEGMTPDLI
1121  FNPHGMPTRMTMGQLAETLASKAAANIGELFDGTPFSDYDVTEIPKVLKELGMDEYGTEEMYCGMTGRKMKARIFIGPMF
1201  YLRLKHMVLDKVHSRSMGPRQAITRQPMEGRSKDGGLKIGEMEKDAMVAHGIGQFLKERMMENSDITTIMVCDKCGTFAT
1281  KVMDKEYYVCNNCNNHTDFSNVAMPYAFKLMVQELTAVNILPRIRAEKIDL
```

Fig. 3. Proteomic peptide coverage of the DNA-dependent RNA polymerase II subunit 2 precursor. The intein element is printed in bold letters, peptides identified in this study are shown in red, and residues after which tryptic cleavage could occur are underlined.

candidates identified in this study. Both CroV and Mimivirus package predicted enzymes for DNA repair (DNA polymerase X and a 3′-5′ exonuclease) and protein modification (Ser/Thr protein kinase, Ser/Thr protein phosphatase, Cys protease), as well as an FtsJ-like RNA methyltransferase and a MPP-superfamily metallophosphatase. Finally, CroV and Mimivirus each package three redox-active proteins, which may be involved in virion morphogenesis, as demonstrated for their poxvirus homologs (Senkevich et al., 2002).

Host cell proteins in the virion proteome

Querying the transcriptome data of the host, *C. roenbergensis* strain E4-10, revealed 60 host proteins that matched peptide spectra from our LC–MS/MS analysis (Supplemental file 2). The most abundant host protein (IntVal=3.73E+09) was a tubular mastigoneme-related protein that is likely derived from host-cell flagella (Fig. 1). Other host proteins with well characterized homologs and IntVals >1E+07 include, in order of decreasing IntVals: Inorganic H+ pyrophosphatase, Ras/Ran-family protein, mastigoneme protein (2×), histidine phosphatase, Na/K-transporting ATPase (4×), dynein heavy chain (2×), REJ domain-containing protein, dynamin-family protein, translation elongation factor Tu, valyl-tRNA synthetase, ATP synthase beta subunit, heat shock 70 kDa protein, and a vacuolar transport chaperone-like protein. These proteins may be contaminants that were not removed during density-gradient ultracentrifugation, e.g. by adhesion of host proteins to CroV virions, but the fact that the IntVal for the most abundant host protein was fifty-fold lower than the IntVal for the most abundant CroV protein (crov342) suggests that host protein contamination in the viral preparation is no more than 2%, by weight. Alternatively, host proteins may be located inside the virion, as has been observed for other large DNA viruses such as poxviruses (Resch et al., 2007), herpesviruses (Loret et al., 2008), or baculoviruses (Wang et al., 2010). The roles of host proteins for CroV infection, if any, remain to be determined.

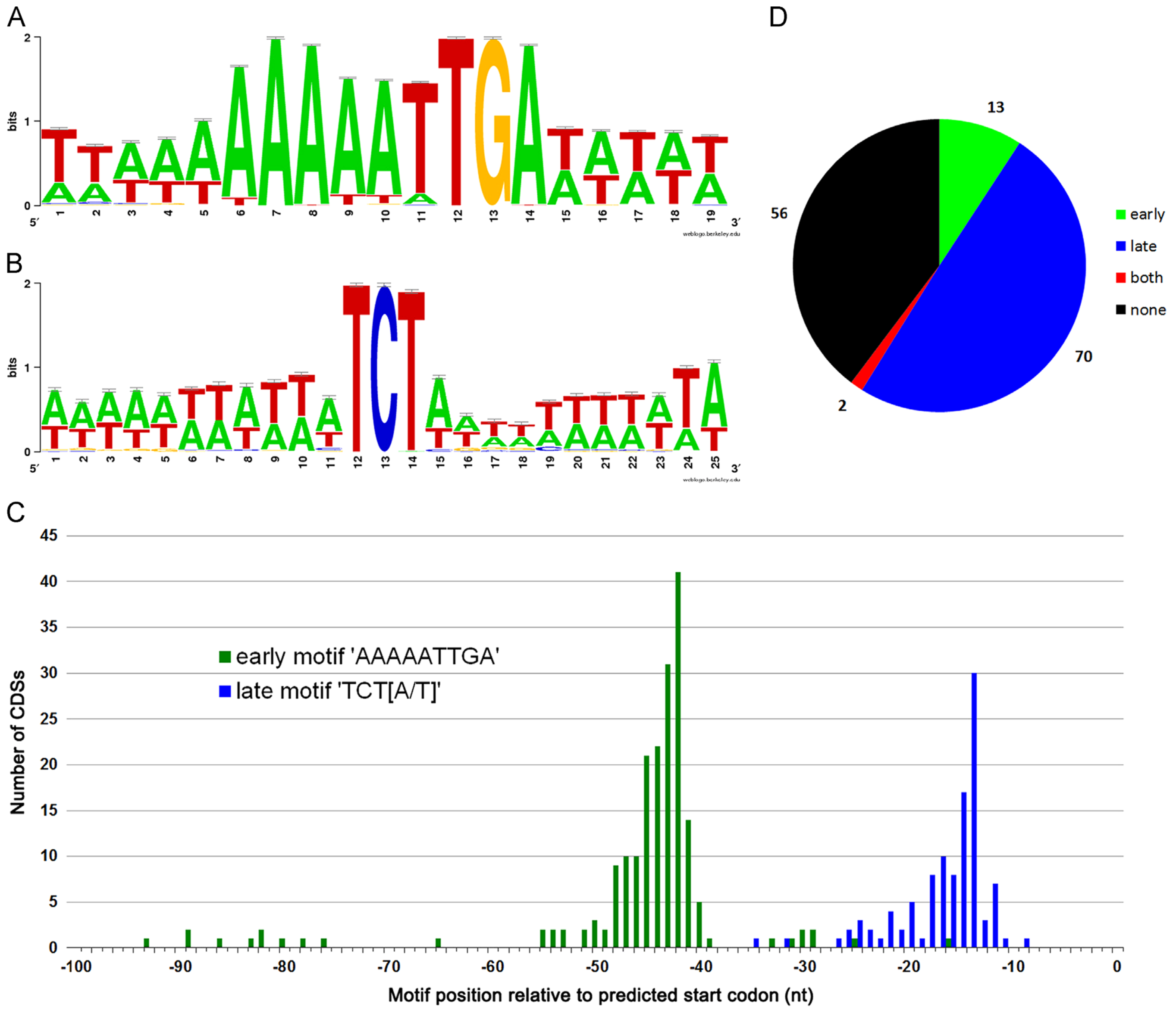


Fig. 4. Updated analysis of gene promoter elements in CroV and their distribution among virion protein-coding genes. (A) Consensus motif of 195 aligned early promoter sequences. (B) Consensus motif of 109 aligned late promoter sequences. (C) Positional distribution of the early and late promoter motifs relative to the predicted start codon. The early and late motifs are preferentially found at positions -42 and -14, respectively, as defined by the number of nucleotides between the first nucleotide of the predicted start codon and the first adenine in 'AAAAATTGA' or the first thymine in 'TCT[A/T]'. (D) Distribution of early and late promoter motifs among the 141 CroV virion protein-coding genes.

With current LC–MS/MS technology it is not possible to fully characterize all proteins in a complex sample such as this; however, given that most of the detected host proteins were on the lower end of the abundance spectrum (Fig. 1), it suggests that most virion proteins were detected.

Conclusion

The particles of giant DNA viruses are highly complex macromolecular assemblies composed of multiple protein and lipid layers, are often endowed with specific portals for genome entry and exit (Klose et al., 2010; Xiao and Rossmann, 2011; Zhang et al., 2011), and may contain upwards of 100 viral proteins (Renesto et al., 2006; Dunigan et al., 2012). In this study, we identified 141 virally encoded proteins in the virion of the giant marine virus CroV, which we believe is close to the complete virion proteome. These proteins were not a random representation of the viral coding potential; instead, the CroV virion proteins could be clustered into several distinct functional categories. Surprisingly, CroV packages a photolyase and enzymes for a presumably complete BER pathway. To our knowledge, this is the most extensive DNA repair machinery found in a virus particle to date. These DNA repair proteins may enable the virus to restore a functional genome, even after exposure to ultraviolet radiation and reactive oxygen species during its extracellular stage. The predicted mechanosensitive channel protein $\mathrm{MscL}_{\mathrm{CroV}}$ could protect the genome-containing core from osmotic damage. The 16 predicted transcription proteins strongly imply that CroV, like Mimivirus and vaccinia virus, employs a cytoplasmic replication strategy that requires the virion presence of a functional transcription complex for early gene expression (Mutsafi et al., 2010). Another conserved feature between poxviruses and members of the *Mimiviridae* is the presence of oxidoreductases and PDIs in the virion, which may be required for the structural integrity of specific virion components. Overall, the composition of the CroV virion proteome and the 36 virion proteins that CroV shares with Mimivirus suggest that protein packaging is an evolutionarily conserved process, and that CDSs crov241–crov282, which are presumed recent acquisitions from a bacterial source, do not contribute towards the virion architecture. Virion proteome analysis of newly discovered giant viruses may thus complement genomic analysis as a useful tool for virus classification and for identification of conserved virion features.

Material and methods

Purification of extracellular CroV virions

C. roenbergensis strain E4-10 was cultured and infected with CroV strain BV-PW1 as described previously (Fischer et al., 2010). Cell concentrations of infected cultures were monitored by staining with Lugol's Acid Iodine solution and counting by light microscopy using a hemocytometer (improved Neubauer counting chamber). 10 l of lysate was centrifuged for 1 h at 10,500 × g in a Sorvall RC-5C centrifuge (GSA rotor, 4 °C) to pellet bacteria (which serve as food source for *C. roenbergensis*) and cell debris. CroV particles were concentrated from the supernatant by tangential flow filtration using a Vivaflow 200, 0.2 µm PES unit (Sartorius Mechatronics, Canada) to a final volume of ~30 ml. The concentrate was loaded in Ultra-clear™ ultracentrifuge tubes (Beckman, Canada), then 0.4 ml of 50% (w/v) Optiprep™ solution (Sigma, Canada) in 50 mM Tris–HCl, pH 7.6 was carefully pipetted to the bottom of the tubes, and the tubes were centrifuged for 1 h at 160,000 × g, 20 °C in a Beckman SW40 rotor (Sorvall RC80 ultracentrifuge). The concentrated layer of viruses and bacteria on top of the Optiprep™ cushion was removed by pipetting and dialyzed overnight in a 20,000 MWCO dialysis cassette (3 ml Slide-A-Lyzer, Pierce, U.S.A.) against 1 l of 200 mM Tris–HCl, pH 7.6, 4 °C, and then for several hours against 1 l of 50 mM Tris–HCl, pH 7.6, 4 °C. In order to separate virus particles from bacterial cells and other contaminating material, the dialyzed suspension was transferred to thin-walled 14 ml Ultra-clear™ ultracentrifuge tubes (Beckman, Canada) on top of a 12 ml 15/25/35% (w/v) Optiprep™ gradient (in 50 mM Tris–HCl, pH 7.6), which had been linearized at room temperature overnight. The gradient was centrifuged for 1.5 h at 111,000 × g, 20 °C in a Beckman SW40 rotor, after which the CroV-containing band (located at ~30% w/v Optiprep™) was extracted by puncturing the side of the tube with a 23 G needle and diluted threefold with 50 mM Tris–HCl, pH 7.6. The purified virions were concentrated by centrifugation for 1 h at 20,000 × g, 20 °C, in 1.5 ml microfuge tubes (Eppendorf 5810 tabletop centrifuge) and the pellets resuspended in 20 µl of 50 mM Tris–HCl, pH 7.6.

Determination of CroV particle concentration

1 µl of the density gradient-purified CroV suspension was mixed with 979 µl of nuclease-free water and 20 µl of 25% (w/v) glutaraldehyde (Sigma-Aldrich Canada). After mixing and incubation at 4 °C for 20 min, the suspension was diluted 10-fold with nuclease-free water for a final dilution factor of 10,000. 800 µl was filtered through a 0.02 µm pore-size membrane filter (Whatman Anodisc, VWR Canada) using a Hoefer ten-place filtration manifold with a 0.45 µm pore-size support filter (HAWP02500, Millipore) and counted by epifluoresence microscopy (Suttle and Fuhrman, 2010). The Anodisc filter was placed topside-up on a 70 µl drop of 0.25% SYBR Green I (Invitrogen Canada) nucleic acid stain and incubated at room temperature in dark for 17 min. To remove excess stain, the Anodisc filter was then placed again on the 0.45 µm pore-size support filter on the filtration manifold and vacuum was applied. The Anodisc filter was mounted on a glass slide using a solution of 0.1% p-phenylenediamine, 50% (v/v) glycerol in phosphate-buffered saline. Virus-like particles on the mounted filter were counted on an epifluorescence microscope (Olympus AX70) at 400 × magnification.

Protein sample fractionation and tryptic digestion

The 20 µl sample containing $\sim 3 \times 10^{10}$ virions was solubilized in SDS buffer with 0.025 U/µl benzonase nuclease (Novagen, Germany) added and allowed to sit for 1 h at 22 °C. Proteins were resolved on a 10% SDS-polyacrylamide gel that was subsequently stained with blue-silver colloidal Coomassie (Candiano et al., 2004). The gel lane was cut into 10 slices and proteins were reduced with 10 mM dithiothreitol for 45 min at 56 °C, alkylated with 55 mM iodoacetamide for 30 min at 37 °C and proteolyzed to peptides with 12.5 ng/µl trypsin for 15 h at 37 °C, as described (Shevchenko et al., 1996).

Liquid chromatography–tandem mass spectrometry

The samples were desalted with a STop-And-Go Extraction (STAGE) tip (Rappsilber et al., 2007) and analyzed by liquid chromatography–tandem mass spectrometry (LC–MS/MS). The LC–MS/MS setup consisted of a linear trapping quadrupole-Orbitrap (LTQ-OrbitrapXL, Thermo Fisher Scientific, Germany), which was coupled on-line to a 1100 Series nanoflow high performance liquid chromatography system (HPLC; Agilent Technologies) with a nanospray ionization source (Proxeon, Denmark) and a 20 cm long and 50 µm inner diameter holding column packed in-house with ReproSil-Pur

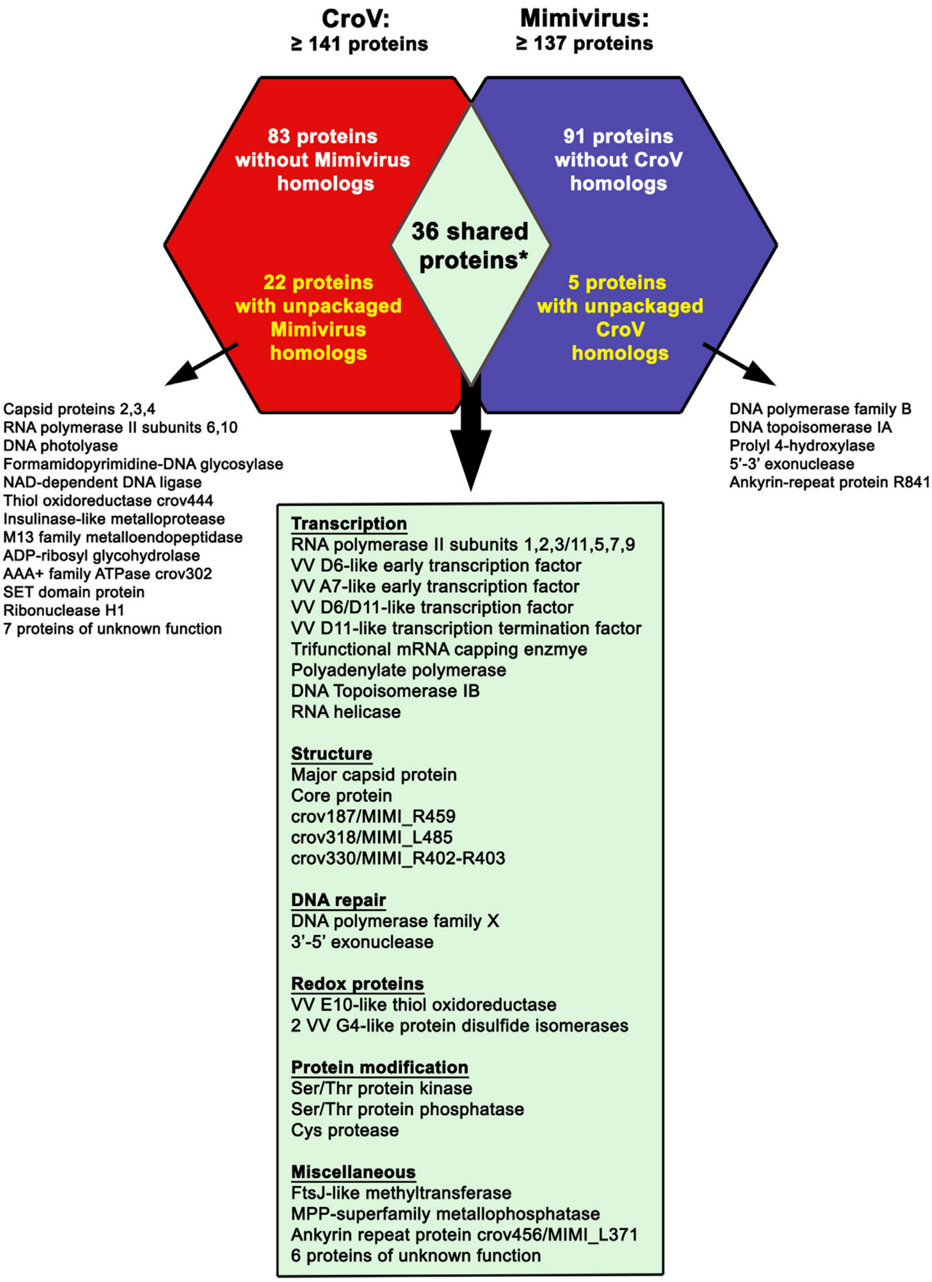

CroV:
≥ 141 proteins
Mimivirus:
≥ 137 proteins
83 proteins without Mimivirus homologs
91 proteins without CroV homologs
36 shared proteins*
22 proteins with unpackaged Mimivirus homologs
5 proteins with unpackaged CroV homologs
Capsid proteins 2,3,4
RNA polymerase II subunits 6,10
DNA photolyase
Formamidopyrimidine-DNA glycosylase
NAD-dependent DNA ligase
Thiol oxidoreductase crov444
Insulinase-like metalloprotease
M13 family metalloendopeptidase
ADP-ribosyl glycohydrolase
AAA+ family ATPase crov302
SET domain protein
Ribonuclease H1
7 proteins of unknown function
DNA polymerase family B
DNA topoisomerase IA
Prolyl 4-hydroxylase
5′-3′ exonuclease
Ankyrin-repeat protein R841
Transcription
RNA polymerase II subunits 1,2,3/11,5,7,9
VV D6-like early transcription factor
VV A7-like early transcription factor
VV D6/D11-like transcription factor
VV D11-like transcription termination factor
Trifunctional mRNA capping enzmye
Polyadenylate polymerase
DNA Topoisomerase IB
RNA helicase
Structure
Major capsid protein
Core protein
crov187/MIMI_R459
crov318/MIMI_L485
crov330/MIMI_R402-R403
DNA repair
DNA polymerase family X
3′-5′ exonuclease
Redox proteins
VV E10-like thiol oxidoreductase
2 VV G4-like protein disulfide isomerases
Protein modification
Ser/Thr protein kinase
Ser/Thr protein phosphatase
Cys protease
Miscellaneous
FtsJ-like methyltransferase
MPP-superfamily metallophosphatase
Ankyrin repeat protein crov456/MIMI_L371
6 proteins of unknown function

Fig. 5. The shared virion proteome of CroV and Mimivirus. Proteins are categorized into confirmed virion components in both viruses, proteins that have homologs in both viruses but were found in the virion of only one virus, and virion proteins found in one virus that lack homologs in the other virus. *Counting CroV proteins.

C18-AQ, 3 µm, 120 A (Dr. Maisch GmBH) (Chan et al., 2006). The trap column contained AQUA C18, 5 µm, 200 A (Phenomenex). Buffer A (0.5% acetic acid) and buffer B (0.5% acetic acid, 80% acetonitrile) were run in gradients of 10–32% B over 57 min, 32–40% B over 5 min, 40–100% B over 2 min, 100% B for 2 min, and finally 0% B for another 10 min to recondition the column (Chan et al., 2006). The LTQ-Orbitrap was set to acquire a full-range scan at 60,000 resolution from 350 to 1500 thomson (Th) in the Orbitrap and to simultaneously fragment the top five peptide ions in each cycle in the LTQ.

Data processing

A custom protein database was first compiled from GenBank and consisted of 544 predicted CroV protein-coding sequences (CDSs, GenBank accession NC_014637), 1352 CroV unidentified reading frames (URFs), 20 Mavirus CDSs (GenBank accession NC_015230), and 53 Mavirus URFs, as well as all predicted proteins from the *C. roenbergensis* transcriptome and the *C. roenbergensis* mitochondrial DNA (GenBank accession NC_000946). MaxQuant (v1.4.0.3) (Cox and Mann, 2008) was used to identify peptides from the fragment spectra using the above database and default parameters. Trypsin was specified as the enzyme, with up to one missed cleavage; cysteine carbamidomethylation was defined as a fixed modification, while methionine oxidation and deamidation of asparagine and glutamine were allowed as variable modifications. A 1% false discovery rate (at the protein level) was used to determine whether a protein was confidently identified. The intensity values were calculated as the average intensity of the five-most intense ions detected for each protein. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (Vizcaíno et al., 2014) via the PRIDE partner repository with the dataset identifier PXD000993. Transmembrane helices in viral proteins were predicted using the Hidden Markov Model implemented in the TMHMM Server v. 2.0 (Krogh et al., 2001). Gene Ontology term enrichment analysis was carried out via the Fisher's exact test (p-value < 0.05) implemented in the Blast2Go suite (Götz et al., 2008). For the CroV–Mimivirus comparative virion proteome analysis, we considered Mimivirus virion proteins that had been reported in the initial study by Renesto et al. (2006) as well as additional proteins identified by Claverie et al. (2009) with an E-value < 0.01.

Acknowledgments

This work was supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grants 311654-11 (to L.J.F) and 109785 (to C.A.S), the Tula Foundation through the Centre for Microbial Diversity and Evolution, and fellowships awarded to M.G.F by the Gottlieb Daimler- and Karl Benz-Foundation, Germany, the University of British Columbia, and the European Molecular Biology Organization (EMBO). Mass spectrometry infrastructure used in this project was supported by the Canada Foundation for Innovation, the British Columbia Knowledge Development Fund and the British Columbia Proteomics Network. L.J.F is the Canada Research Chair in Quantitative Proteomics. Data deposition to the ProteomeXchange Consortium was supported by the PRIDE Team of EMBL-EBI.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.05.029.

References

Allen, M.J., Howard, J.A., Lilley, K.S., Wilson, W.H., 2008. Proteomic analysis of the EhV-86 virion. Proteome Sci. 6, 1–6.

Appenzeller-Herzog, C., Ellgaard, L., 2008. The human PDI family: versatility packed into a single fold. Biochim. Biophys. Acta 1783, 535–548.

Booth, I.R., Blount, P., 2012. The MscS and MscL families of mechanosensitive channels act as microbial emergency release valves. J. Bacteriol. 194, 4802–4809.

Broyles, S.S., 2003. Vaccinia virus transcription. J. Gen. Virol. 84, 2293–2303.

Candiano, G., Bruschi, M., Musante, L., Santucci, L., Ghiggeri, G.M., Carnemolla, B., Orecchia, P., Zardi, L., Righetti, P.G., 2004. Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. Electrophoresis 25, 1327–1333.

Chan, Q.W.T., Howes, C.G., Foster, L.J., 2006. Quantitative comparison of caste differences in honeybee hemolymph. Mol. Cell Proteomics 5, 2252–2262.

Chung, C., Chen, C., Ho, M., Huang, C., 2006. Vaccinia virus proteome: identification of proteins in vaccinia virus intracellular mature virion particles. J. Virol. 80, 2127–2140.

Claverie, J.-M., Abergel, C., Ogata, H., 2009. Mimivirus. Curr. Top. Microbiol. Immunol. 328, 89–121.

Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X.-W., Federici, B.a., Van Etten, J.L., Koonin, E.V., La Scola, B., Raoult, D., 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch. Virol. 158, 2517–2521.

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372.

Dunigan, D.D., Cerny, R.L., Bauman, A.T., Roach, J.C., Lane, L.C., Agarkova, I.V., Wulser, K., Yanai-Balser, G.M., Gurnon, J.R., Vitek, J.C., Kronschnabel, B.J., Jeanniard, A., Blanc, G., Upton, C., Duncan, G.A., McClung, O.W., Ma, F., Van Etten, J.L., 2012. Paramecium bursaria chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. J. Virol. 86, 8821–8834.

Fenchel, T., Patterson, D.J., 1988. *Cafeteria roenbergensis* nov. gen., nov. sp., a heterotrophic microflagellate from marine plankton. Mar. Microb Food Webs 3, 9–19.

Fischer, M.G., Suttle, C.A., 2011. A virophage at the origin of large DNA transposons. Science 332, 231–234.

Fischer, M.G., Allen, M.J., Wilson, W.H., Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc. Natl. Acad. Sci. USA 107, 19508–19513.

Garza, D.R., Suttle, C.A., 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.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435.

Iyer, L.M., Aravind, L., Koonin, E.V., 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. 75, 11720–11734.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol (in press).

Klose, T., Kuznetsov, Y.G., Xiao, C., Sun, S., McPherson, A., Rossmann, M.G., 2010. The three-dimensional structure of mimivirus. Intervirology 53, 268–273.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

Loret, S., Guay, G., Lippé, R., 2008. Comprehensive characterization of extracellular herpes simplex virus type 1 virions. J. Virol. 82, 8605–8618.

Mackinder, L.C.M., Worthy, C.A., Biggi, G., Hall, M., Ryan, K.P., Varsani, A., Harper, G.M., Wilson, W.H., Brownlee, C., Schroeder, D.C., 2009. A unicellular algal virus, *Emiliania huxleyi* virus 86, exploits an animal-like infection strategy. J. Gen. Virol. 90, 2306–2316.

Mutsafi, Y., Zauberman, N., Sabanay, I., Minsky, A., 2010. Vaccinia-like cytoplasmic replication of the giant Mimivirus. Proc. Natl. Acad. Sci. USA 107, 5978–5982.

O'Day, D.H., Suhre, K., Myre, M.A., Chatterjee-Chakraborty, M., Chavez, S.E., 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.

Randall, A.Z., Baldi, P., Villarreal, L.P., 2004. Structural proteomics of the poxvirus family. Artif. Intell. Med. 31, 105–115.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350.

Rappsilber, J., Mann, M., Ishihama, Y., 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. 2, 1896–1906.

Redrejo-Rodríguez, M., Salas, M.L., 2014. Repair of base damage and genome maintenance in the nucleo-cytoplasmic large DNA viruses. Virus Res. 179, 12–25.

Renesto, P., Abergel, C., Decloquement, P., Moinier, D., Azza, S., Ogata, H., Fourquet, P., Gorvel, J.P., Claverie, J.-M., Raoult, D., 2006. Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J. Virol. 80, 11678–11685.

Resch, W., Hixson, K.K., Moore, R.J., Lipton, M.S., Moss, B., 2007. Protein composition of the vaccinia virus mature virion. Virology 358, 233–247.

Ryser, H.J., Levy, E.M., Mandel, R., DiSciullo, G.J., 1994. Inhibition of human immunodeficiency virus infection by agents that interfere with thiol-disulfide interchange upon virus-receptor interaction. Proc. Natl. Acad. Sci. USA 91, 4559–4563.

Schelhaas, M., Malmström, J., Pelkmans, L., Haugstetter, J., Ellgaard, L., Grünewald, K., Helenius, A., 2007. Simian Virus 40 depends on ER protein folding and quality control factors for entry into host cells. Cell 131, 516–529.

Senkevich, T.G., White, C.L., Koonin, E.V., Moss, B., 2002. Complete pathway for protein disulfide bond formation encoded by poxviruses. Proc. Natl. Acad. Sci. USA 99, 6667–6672.

Shevchenko, A., Wilm, M., Vorm, O., Mann, M., 1996. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal. Chem. 68, 850–858.

Song, W., Lin, Q., Joshi, S.B., Lim, T.K., Hew, C.-L., 2006. Proteomic studies of the Singapore grouper iridovirus. Mol. Cell Proteomics 5, 256–264.

Srinivasan, V., Tripathy, D.N., 2005. The DNA repair enzyme, CPD-photolyase restores the infectivity of UV-damaged fowlpox virus isolated from infected scabs of chickens. Vet. Microbiol. 108, 215–223.

Suhre, K., Audic, S., Claverie, J.-M., 2005. Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. Proc. Natl. Acad. Sci. USA 102, 14689–14693.

Suttle, C.A., Fuhrman, J.A., 2010. Enumeration of virus particles in aquatic or sediment samples by epifluorescence microscopy. In: Wilhelm, S., Weinbauer, M., Suttle, C. (Eds.), Manual of Aquatic Viral Ecology. American Society of Limnology and Oceanography, Waco, TX, pp. 145–153.

Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R.J., Kraus, H.-J., Albar, J.P., Martinez-Bartolomé, S., Apweiler, R., Omenn, G.S., Martens, L., Jones, A.R., Hermjakob, H., 2014.

ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32, 223–226.

Wang, R., Deng, F., Hou, D., Zhao, Y., Guo, L., Wang, H., Hu, Z., 2010. Proteomics of the Autographa californica nucleopolyhedrovirus budded virions. J. Virol. 84, 7233–7242.

Xiao, C., Rossmann, M.G., 2011. Structures of giant icosahedral eukaryotic dsDNA viruses. Curr. Opin. Virol. 1, 101–109.

Yoder, J.D., Chen, T.S., Gagnier, C.R., Vemulapalli, S., Maier, C.S., Hruby, D.E., 2006. Pox proteomics: mass spectrometry analysis and identification of Vaccinia virion proteins. Virol. J. 3, 10.

Yutin, N., Koonin, E.V., 2012. Hidden evolutionary complexity of nucleo-cytoplasmic large DNA viruses of eukaryotes. Virol. J. 9, 161.

Zhang, X., Xiang, Y., Dunigan, D.D., Klose, T., Chipman, P.R., Van Etten, J.L., Rossmann, M.G., 2011. Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. Proc. Natl. Acad. Sci. USA 108, 14837–14842.