GENETIC AND ULTRASTRUCTURAL CHARACTERIZATION OF *CAFETERIA ROENBERGENSIS* VIRUS AND ITS VIROPHAGE MAVIRUS

by

MATTHIAS GUNTHER FISCHER

Dipl. Biochem., University of Bayreuth, Germany, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Microbiology and Immunology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

March 2011

© Matthias Gunther Fischer, 2011

Abstract

Giant viruses infecting unicellular eukaryotes have genomes that overlap in size and coding content with the smallest cellular life forms, thereby blurring the boundary between what is considered living and non-living. Due to their recent discovery, little is known about the biology and host range of giant viruses. In this dissertation, I characterize *Cafeteria roenbergensis* virus (CroV), the largest marine virus known to date. CroV infects the phagotrophic nanoflagellate *C. roenbergensis*, a widespread and ecologically important marine zooplankton species. CroV has a 730 kilobase pair DNA genome which is predicted to encode 544 proteins and 22 transfer RNAs. Four genes contained an intein insertion and several genes have not been found before in viruses, including an isoleucyl-tRNA synthetase and a histone acetyltransferase. A 38 kilobase pair region of putative bacterial origin encoded predicted enzymes for the biosynthesis of 3-deoxy-D-*manno*-octulosonate, a key component of the bacterial lipopolysaccharide layer. Microarray analysis revealed that at least 274 CroV genes were transcribed during infection and that different genes were expressed at early and late stages of viral replication. Promoter sequences specific for each stage were identified. Proteomic analyses showed that the virion is composed of at least 129 CroV-encoded proteins, including a large set of transcription enzymes and several DNA repair proteins. Phylogenetically, CroV was found to belong to the group of nucleocytoplasmic large DNA viruses and was most closely related to *Acanthamoeba polyphaga* mimivirus, although only a third of the CroV genes had homologues in Mimivirus. I also discovered a smaller virus, the Mavirus virophage, whose replication was dependent on co-infection by CroV and led to decreased CroV production and increased host-cell survival. Mavirus particles co-localized within the CroV virion factory, as shown by transmission electron microscopy of infected cells. Remarkably, the 19 kilobase pair DNA genome of Mavirus was most similar to the *Maverick/Polinton* eukaryotic DNA transposons, which led to the hypothesis that these transposons have originated from the endogenization of ancient virophages into eukaryotic genomes. This work describes the first giant virus infecting a zooplankton species and demonstrates a clear link between Mavirus and *Maverick* transposons.

Preface

The work presented in this dissertation would not have been possible without contributions made by collaborators, contractors, and undergraduate students, as described in the following paragraphs. In chapters 2 and 4, 454 pyrosequencing and automated contig assembly was contracted at 454 Life Sciences, Branford, CT, USA, and at the McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada. The amino acid alignment of isoleucyl-tRNA synthetases, on which the phylogenetic tree in Fig. 2.9 is based, was constructed by Shi 'Small' Mang as part of an undergraduate research project.

In chapter 2, the microarray study was conceived by Drs. Michael Allen and William Wilson at the Plymouth Marine Laboratories, Plymouth, UK, and was carried out in the laboratory of Dr. Wilson under the supervision of Dr. Allen. Dr. Allen designed the oligonucleotides and ordered the microarray slides, which were manufactured at the Liverpool Microarray Facility, UK. He also made significant contributions to experimentation and data analysis and was responsible for deposition of the microarray data in the GEO database. For microarray analysis, I planned and conducted the infection experiments in *Cafeteria roenbergensis*, extracted mRNA from infected cells, performed the microarray experiments (cDNA synthesis and labelling, hybridization, image analysis) under the guidance of Dr. Allen, and analyzed the microarray data.

In chapter 3, the proteome study was conceived by Drs. Leonard Foster and Curtis Suttle of the University of British Columbia. SDS-PAGE, enzymatic treatments, LC-MS/MS analysis, and mass spec data analysis were conducted by Dr. Isabelle Kelly in the laboratory of Dr. Foster, who contributed to data analysis and manuscript revision. I provided pure virions for mass spectrometry, analyzed the data, and wrote the manuscript, for which Dr. Kelly contributed part of the materials and methods section.

In chapter 5, Dr. Garnet Martens and Bradford Ross from the UBC Bioimaging Facility were contracted to perform cryo-fixation, freeze-substitution, osmication, block preparation, ultra-thin sectioning, and staining of TEM samples. Amy Chan assisted during preparation of infected cells for cryo-fixation. I conceived and performed the infection experiments, collected and analyzed the TEM data, and wrote the manuscript.

In chapters 1-6, I performed all other experimental work, data analysis, and manuscript writing. My contributions included maintenance of *C. roenbergensis* cultures, performance of infection studies, virus purification and quantification, and gap-closing of the CroV genome, which

involved the construction of genomic shotgun libraries as well as PCR experiments and bioinformatic methods. Furthermore, I was responsible for annotation of the genomes of CroV, Mavirus, and phage307. I discovered the Mavirus virophage and developed the hypothesis of viral transposogenesis for *Maverick/Polinton* DNA transposons.

As research advisor, Dr. Curtis Suttle was involved in all aspects of this work, including conceptualization of the experiments and manuscript writing.

A version of chapter 2 has been published: Fischer M.G., Allen M.J., Wilson W.H. and Suttle C.A. (2010) A giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences USA,* Vol. 107, pp. 19508-19513. A version of chapter 4 has been published: Fischer M.G. and Suttle C.A. (2011) A virophage at the origin of large DNA transposons. *Science*, March 3, DOI: 10.1126/science.1199412. Manuscripts of chapters 3 and 5 are in preparation for submission to peer-reviewed journals.

This work has been approved by the University of British Columbia's Biosafety Board (Biosafety certificate B10-0059-001 DG2010-2015).

Table of Contents

Abstract ..... **ii**

Preface ..... **iii**

Table of Contents ..... **v**

List of Tables ..... **x**

List of Figures ..... **xi**

List of Abbreviations ..... **xiv**

Acknowledgements ..... **xix**

Dedication ..... **xx**

1 Introduction to Large Aquatic DNA Viruses **1**

1.1 Viruses in the marine environment 1

1.2 Giant DNA viruses 1

1.2.1 The nucleocytoplasmic large DNA viruses 3

1.2.2 Mimivirus 5

1.2.3 How common are giant viruses in the environment? 7

1.3 Virophages 9

1.3.1 The Sputnik virophage 9

1.3.2 Virophages and other satellite viruses 10

1.4 Origin and evolution of viruses 11

1.5 *Cafeteria roenbergensis* virus 13

1.5.1 *Cafeteria* – a genus of heterotrophic nanoflagellates 13

1.5.2 Viruses infecting *Cafeteria roenbergensis* 15

1.6 Thesis objectives 16

2 Genome Analysis of CroV **17**

2.1 Synopsis 17

2.2 Materials and methods 17

2.2.1 Flagellate growth and virus purification 17
2.2.2 Pulsed-field gel electrophoresis of CroV genomic DNA 19
2.2.3 Genome sequencing and assembly 20
2.2.4 Genome annotation 21
2.2.5 Phylogenetic analysis 22
2.2.6 Microarray analysis 23
2.3 Results and discussion 27
2.3.1 General genome features 27
2.3.2 Translation genes 35
2.3.3 DNA repair genes 38
2.3.4 Transcription genes 44
2.3.5 DNA replication enzymes 46
2.3.6 Ubiquitin pathway components 47
2.3.7 Molecular chaperones and redox proteins 48
2.3.8 Inteins 52
2.3.9 Repetitive DNA elements 55
2.3.10 Microarray analysis 58
2.3.11 Promoter analysis 60
2.3.12 A 38 kbp genomic fragment involved in carbohydrate metabolism 62
2.3.13 Phylogenetic relationship to other viruses 70
2.4 Conclusions 74
3 The Virion Proteome of CroV 75
3.1 Synopsis 75
3.2 Materials and methods 75
3.2.1 Virus purification 75
3.2.2 Protein sample fractionation and tryptic digestion 76
3.2.3 Liquid chromatography-tandem mass spectrometry 76
3.2.4 Data processing 78

3.3 Results and discussion....79
3.3.1 Identification of CroV virion proteins....79
3.3.2 Structural proteins....90
3.3.3 Transcription components....91
3.3.4 DNA repair enzymes....92
3.3.5 Redox proteins....93
3.3.6 Comparison with microarray data....95
3.3.7 Comparison with CroV promoters....96
3.3.8 CroV and Mimivirus share a conserved set of virion proteins....97
3.3.9 Proteins from Mavirus and phage307....98
3.4 Conclusions....100
4 Mavirus, a Novel Virophage Associated with CroV....102
4.1 Synopsis....102
4.2 Materials and methods....102
4.2.1 Culture conditions and virophage purification....102
4.2.2 Genome sequencing and annotation....103
4.2.3 Co-infection experiments....104
4.2.4 Quantitative PCR....104
4.3 Results and discussion....106
4.3.1 Mavirus replication in *Cafeteria roenbergensis*....106
4.3.2 The Mavirus genome....107
4.3.3 Comparison of Mavirus and Sputnik virophages....117
4.3.4 Mavirus is related to *Maverick/Polinton* DNA transposons....120
4.3.5 Viral transposogenesis of *Maverick/Polinton* DNA transposons....122
4.4 Conclusions....125
5 Ultrastructural Characterization of CroV and Mavirus Infection....126
5.1 Synopsis....126
5.2 Materials and methods....126

5.2.1 Transmission electron microscopy 126
5.2.2 Estimation of burst size 127
5.3 Results and discussion 128
5.3.1 Ultrastructure of *Cafeteria roenbergensis* 128
5.3.2 CroV cell entry and uncoating 130
5.3.3 The CroV virion factory 132
5.3.4 The CroV virion and cell lysis 137
5.3.5 CroV infection may be supported by ribosome aggregates 139
5.3.6 CroV burst size 141
5.3.7 Clathrin-mediated endocytosis of Mavirus 141
5.3.8 Mavirus replication within the CroV virion factory 144
5.4 Conclusions 149
6 Concluding Chapter 151
6.1 The infection cycles of CroV and Mavirus 151
6.1.1 Before cell entry: the CroV particle 152
6.1.2 Cell entry of CroV 153
6.1.3 Early steps in CroV infection 153
6.1.4 CroV capsid assembly, genome packaging, and cell lysis 155
6.1.5 The infection cycle of the Mavirus virophage 155
6.2 CroV genome evolution 157
6.2.1 The origin of CroV genes 157
6.2.2 Indicators for an ancestral origin of CroV 159
6.3 Taxonomic classification 160
6.4 Limitations and outlooks 161
6.5 Significance of the work 164
Bibliography 167
Appendices 190
Appendix A: 454 pyrosequencing and contig assembly 190

Appendix B: Annotation of CroV CDSs 198
Appendix C: Physical properties of CroV CDSs 243
Appendix D: List of PCR primers 260
Appendix E: The genome of phage307 273
Appendix F: CroV in the media 278

List of Tables

Table 1.1 Giant viruses. 2

Table 1.2 Family characteristics of nucleocytoplasmic large DNA viruses. 5

Table 2.1 Analysis of the PFGE restriction patterns from Fig. 2.4. 30

Table 2.2 Novel viral features in the CroV genome. 34

Table 2.3 Transfer RNA genes in the CroV genome. 37

Table 2.4 Properties of the four CroV inteins. 54

Table 2.5 Top BLASTP results for the 34 CDSs in the 38 kbp genomic fragment. 64

Table 3.1 Viral proteins identified in this study through LC-MS/MS analysis. 81

Table 3.2 Functional categories represented by CroV virion proteins. 89

Table 4.1 Cell and virus concentrations during infection of *C. roenbergensis* with CroV and Mavirus. 107

Table 4.2 Quantitative PCR analysis of Mavirus replication. 107

Table 4.3 Mavirus gene annotations. 112

Table 4.4 Physical properties of Mavirus CDSs. 116

Table 4.5 Sequence similarities of Mavirus and Sputnik proteins as determined by BLASTP analysis. 119

Table A.1 Contig statistics and location of contigs in the assembled CroV genome. 193

Table B.1 Functional annotations of CroV CDSs. 198

Table C.1 Physical properties of CroV CDSs 243

Table D.1 Oligonucleotide primers used in this study. 260

Table E.1 Putative CDSs of phage307 and their annotation. 275

List of Figures

Figure 1.1 The heterotrophic nanoflagellate *Cafeteria roenbergensis* strain E4-10. 14

Figure 2.1 Denaturation of CroV particles.....18

Figure 2.2 Fluorescence intensity profile and threshold settings for a typical microarray hybridization.....26

Figure 2.3 Pulsed-field gel electrophoresis of CroV genomic DNA. 28

Figure 2.4 Pulsed-field gel electrophoresis of CroV restriction digests.....29

Figure 2.5 Genome diagram of CroV.....31

Figure 2.6 Codon usage in CroV. 32

Figure 2.7 Functional categories of Clusters of Orthologous Groups of proteins (COGs) identified in CroV. 33

Figure 2.8 Distribution of top BLASTP hits for CroV CDSs.....33

Figure 2.9 Phylogenetic analysis of isoleucyl-tRNA synthetases. 36

Figure 2.10 The CroV tRNA gene cluster. 37

Figure 2.11 Phylogenetic analysis of AP endonucleases. 39

Figure 2.12 Phylogenetic analysis of MutS ATPases. 40

Figure 2.13 Phylogenetic analysis of the photolyase/chryptochrome family.....41

Figure 2.14 Phylogenetic analysis of CPD class I photolyases. 42

Figure 2.15 Phylogenetic analysis of the predicted DNA photolyase crov149.....43

Figure 2.16 Phylogenetic analysis of ELP3-like histone acetyltransferases. 45

Figure 2.17 Phylogenetic analysis of type IB DNA topoisomerases. 46

Figure 2.18 Phylogenetic analysis of HSP70 proteins. 49

Figure 2.19 Phylogenetic analysis of ATP-dependent Lon proteases. 51

Figure 2.20 Location of CroV inteins in their host proteins. 53

Figure 2.21 Dot plot analysis of the CroV genome to show repeated sequences. 56

Figure 2.22 Sequence logos of FNIP/IP22 repeats in CroV. 57

Figure 2.23 Temporal CroV gene expression profile during infection.....59

Figure 2.24 Early and late gene promoter motifs in CroV.....61

Figure 2.25 MEME sequence logo of the late promoter motif. 61

Figure 2.26 The predicted 3-deoxy-D-manno-octulosonate (KDO) biosynthesis pathway in CroV. 63

Figure 2.27 Phylogenetic analysis of arabinose-5-phosphate isomerases. 66

Figure 2.28 Phylogenetic analysis of KDO 8-P synthases. 67

Figure 2.29 Phylogenetic analysis of KDO 8-P phosphatases. 68

Figure 2.30 Phylogenetic analysis of two types of cytidylyltransferases. 69

Figure 2.31 NCLDV core genes found in CroV. 71

Figure 2.32 Phylogenetic reconstruction of NCLDV members based on DNA polymerase B. 72

Figure 2.33 Phylogenetic analysis of the NCLDV major capsid protein. 73

Figure 3.1 Approximate genomic distribution of genes encoding virion proteins 80

Figure 3.2 Proteomic peptide coverage of the DNA-dependent RNA polymerase II subunit 2 precursor. 92

Figure 3.3 Sequence alignment of Erv1/Alr-family thiol oxidoreductases. 94

Figure 3.4 Sequence alignment of thioredoxins related to the vaccinia virus G4 protein. 94

Figure 3.5 Time-resolved transcriptome data of CroV virion proteins. 96

Figure 3.6 The shared virion proteome of CroV and Mimivirus. 99

Figure 4.1 The Mavirus genome. 108

Figure 4.2 The Mavirus promoter motif. 109

Figure 4.3 Phylogenetic tree of protein-primed DNA-dependent DNA polymerase B. 110

Figure 4.4 Alignment of DNA polymerase B inteins from Mimivirus and CroV, and potential mini-inteins from *Polysphondylium pallidum* and Mavirus. 113

Figure 4.5 Phylogenetic tree of FtsK-HerA-like DNA packaging ATPases. 114

Figure 4.6 Phylogenetic tree of cysteine proteases. 115

Figure 4.7 Gene homologies between Mavirus and Sputnik 117

Figure 4.8 Multiple sequence alignments of virophage and MP transposon capsid proteins. 118

Figure 4.9 Phylogenetic position of the Mavirus retroviral integrase. 121

Figure 4.10 Syntenic regions of Mavirus and an MP-like element from *Polysphondylium pallidum*.. 123

Figure 4.11 Hypothesis for the origin of the *Maverick/Polinton* class of DNA transposons. 124
Figure 5.1 Ultrastructure of *Cafeteria roenbergensis* strain E4-10 129
Figure 5.2 Cytoplasmic CroV cores during early infection. 131
Figure 5.3 Different stages during expansion of the CroV virion factory.....133
Figure 5.4 Ultrastructural aspects of CroV capsid assembly. 134
Figure 5.5 Mature CroV virion factory and vesicles bearing empty capsids. 135
Figure 5.6 Ultrastructural details of CroV particle filling. 136
Figure 5.7 Mature intracellular CroV virions.....138
Figure 5.8 Late stages of CroV infection.....139
Figure 5.9 Ribosome aggregates in *C. roenbergensis*. 140
Figure 5.10 Clathrin-mediated endocytosis of Mavirus. 143
Figure 5.11 Mavirus particles co-localize with the CroV virion factory. 145
Figure 5.12 Intracellular virions of CroV and Mavirus. 146
Figure 5.13 Paracrystalline arrays of Mavirus and defective CroV capsid structures 148
Figure 6.1 The infection cycles of CroV and Mavirus in *C. roenbergensis*. 152
Figure 6.2 Giant viruses and their hosts on the tree of eukaryotes. 164
Figure A.1 Assembly map of the CroV genome showing the locations of 454 contigs. 192
Figure E.1 Consensus sequence of the terminal direct repeat in phage307. 273
Figure E.2 Genome diagram of phage307.....274

List of Abbreviations

aa ..... Amino acid
AcMNPV ..... *Autographa californica* multicapsid nucleopolyhedrovirus
ACMV ..... *Acanthamoeba castellanii* mamavirus
amu ..... Atomic mass unit
AMV ..... *Amsacta moorei* entomopoxvirus
AP ..... Apurinic/apyrimidinic
API ..... Arabinose-5-phosphate isomerase
APMV ..... *Acanthamoeba polyphaga* mimivirus
ASFV ..... African swine fever virus
ATCV ..... *Acanthocystis turfacea* virus
ATP ..... Adenosine-5'-triphosphate
BER ..... Base excision repair
BI ..... Bayesian inference
BLAST ..... Basic local alignment search tool
bp ..... Base pair
BSA ..... Bovine serum albumin
BV-PW1 ..... *Bodo* virus pier water isolate 1
cDNA ..... Complementary DNA
CDS ..... Protein-coding sequence
CEO ..... Capsid-encoding organism
CeV ..... *Chrysochromulina ericina* virus
CIV ..... Chilo iridescent virus
CKS ..... Cytidine monophosphate-KDO synthetase
CME ..... Clathrin-mediated endocytosis
CMP-NeuAcS ..... *N*-acylneuraminate cytidylyltransferase
COG ..... Clusters of orthologous groups of proteins
CPD ..... Cyclobutane pyrimidine dimer
CroV ..... *Cafeteria roenbergensis* virus
DMKMT ..... Demethylmenaquinone methyltransferase
DNase ..... Deoxyribonuclease
dNTP ..... Deoxyribonucleotide triphosphate
DOD ..... Dodecapeptide motif family of homing endonucleases
DpAV4 ..... *Diadromus pulchellus* ascovirus 4a

dsDNA ..... Double-stranded deoxyribonucleic acid
DTT ..... Dithiothreitol
EDTA ..... Ethylenediaminetetraacetic acid
EhV ..... *Emiliania huxleyi* virus
eIF ..... Eukaryotic translation initiation factor
Elp3 ..... Elongator complex protein 3
ER ..... Endoplasmic reticulum
ERV ..... Essential for respiration and viability
EsV ..... *Ectocarpus siliculosus* virus
E-value ..... Expectation value
FirrV ..... *Feldmannia irregularis* virus
FNIP/IP22 ..... [Single-letter code for conserved amino acids of the 22 aa long repeat]
FPV ..... Fowlpox virus
FV3 ..... Frog virus 3
GOS ..... Global ocean sampling
GS ..... Genome sequencer
HAT ..... Histone acetyl transferase
HaV ..... *Heterosigma akashiwo* virus
HcDNAV ..... *Heterocapsa circularisquama* DNA virus
HGT ..... Horizontal gene transfer
HPLC ..... High performance liquid chromatography
HSP ..... Heat shock protein
HvAV3 ..... *Heliothis virescens* ascovirus 3e
IIV ..... Invertebrate iridescent virus
IleRS ..... Isoleucyl-tRNA synthetase
INT ..... Integrase
IS ..... Insertion sequence
ISNKV ..... Infectious spleen and kidney necrosis virus
JRC ..... Jelly-roll capsid
kb ..... Kilobase
KDO 3-deoxy-D-*manno*-octulosonate
KDOPase ..... 3-deoxy-D-*manno*-octulosonate 8-phosphate phosphatase
KDOPS ..... 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase
LC-MS/MS ..... Liquid chromatography-tandem mass spectrometry
LDV ..... Lymphocystis disease virus
LRR ..... Leucine-rich repeat

LTQ ..... Linear trapping quadrupole
LUCA ..... Last universal common ancestor
M ..... Molar
MALDI ..... Matrix-assisted laser desorption ionization
Mavirus ..... *Maverick* virus
MCP ..... Major capsid protein
MCV ..... *Molluscum contagiosum* virus
MEME ..... Multiple expectation maximation algorithm for motif elicitation
MIAME ..... Minimum information about a microarray experiment
ML ..... Maximum likelihood
MOI ..... Multiplicity of infection
MP ..... *Maverick/Polinton*
MpV ..... *Micromonas pusilla* virus
MSV ..... *Melanoplus sanguinipes* entomopoxvirus
MUSCLE ..... Multiple sequence comparison by log-expectation
MV ..... Marseillevirus
NAD ..... Nicotinamide adenine dinucleotide
NCBI ..... National Center for Biotechnology Information
NCLDV ..... Nucleo-cytoplasmic large DNA viruses
NCVOGS ..... Nucleo-cytoplasmic virus orthologous genes
nt ..... Nucleotide
ORF ..... Open reading frame
OtV ..... *Ostreococcus tauri* virus
p.i. ..... Post infection
PAGE ..... Polyacrylamide gel electrophoresis
PBCV ..... *Paramecium bursaria* chlorella virus
PBS ..... Phosphate buffered saline
PCR ..... Polymerase chain reaction
PDI ..... Protein disulfide isomerase
PEP ..... Phosphoenole pyruvate
Pfam ..... Protein families
PFGE ..... Pulsed-field gel electrophoresis
P_i ..... Inorganic phosphate
pI ..... Isoelectric point
PMSF ..... Phenylmethylsulfonyl fluoride
PolB ..... DNA-dependent DNA polymerase B

PoV ..... *Pyramimonas orientalis* virus
PP_i ..... Pyrophosphate
ppt ..... Parts per thousand
PpV ..... *Phaeocystis pouchetii* virus
Q-TOF ..... Quadrupole time-of-flight
RA ..... Relative abundance
rDNA ..... Ribosomal DNA
REO ..... Ribosome-encoding organism
REV ..... Reticuloendotheliosis virus
RNase ..... Ribonuclease
RNR ..... Ribonucleotide reductase
RPB ..... DNA-dependent RNA polymerase
rve-INT ..... Retroviral integrase
S3H ..... Superfamily 3 helicase
SDS ..... Sodium dodecyl sulphate
SSC ..... Saline-sodium citrate
SUI ..... Suppressor of initiator codon mutations
TBE ..... Tris-borate-EDTA
TDP-DHMT dTDP-6-deoxy-L-hexose 3-O-methyltransferase
TE ..... Transposable element or Tris-EDTA
TEM ..... Transmission electron microscopy
TF ..... Transcription factor
TFF ..... Tangential flow filtration
Th ..... Thomson (SI unit of mass-to-charge ratio, 1 Th = 1 Da/e^-)
TM ..... Transmembrane
TMH ..... Transmembrane helix
TMHMM ..... Transmembrane prediction using Hidden Markov Models
TnAV2 ..... *Trichoplusia ni* ascovirus 2c
Topo ..... DNA topoisomerase
TPR ..... Tetratrico peptide repeat
tRNA ..... Transfer ribonucleic acid
TRX ..... Thioredoxin
Ub ..... Ubiquitin
URF ..... Unidentified reading frame
UTR ..... Untranslated region
VF ..... Virion factory

VHG ..... Viral hallmark gene
VLP ..... Virus-like particles
VV ..... Vaccinia virus
XPG ..... Xeroderma pigmentosum

Acknowledgements

My deepest thanks and appreciation go to my wonderful family: my parents who supported me and my decisions throughout the years and allowed me to pursue my dreams, even if it meant putting a distance of 8,360 km between us; and my lovely wife Simone, who left her own family behind to be part of this journey. I sincerely thank my advisor Curtis Suttle for providing me the opportunity to work on this exciting project and for his guidance, expertise, and support. My thanks go to all the members of the Suttle laboratory, past and present - you have enriched my life by scientific discussions, entertainment, and leisurely activities! In particular, I want to thank Alex Culley, Amy Chan, André Comeau, Caroline Chénard, Christina Charlesworth, Danielle Winget, Emma Hambly, Emma Shelford, Jan Finke, Jérôme Payet, Jessica Labonté, Jessie Clasen, Johan Vande Voorde, Julia Gustavson, Marli Vlok, and Renat Adelshin. Thanks also to my former undergraduate students John Li, Small Mang, and Tyler Nelson for putting so much effort and enthusiasm into their projects. Chris, Dany, and Paul – thank you for many enjoyable rounds of disc golf, rides on Whistler/Blackcomb, and rolls of Sushi! Next, I would like to thank Steven Hallam, François Jean, Brian Leander, and Curtis Suttle for serving on my PhD advisory committee and for their support and advice. I am grateful to Chuan Xiao, Walter Karlen, Tom Beatty, Julia Gustavson, Kersten Rabe, Jessica Labonté, Danielle Winget, Caroline Chénard, , Eric Jan, Michael Allen, and Ellen Pritham for proofreading and valuable comments on manuscripts. I also want to thank Roger Hendrix for his help with phage307 genome annotation. Too numerous to list are the people whom I owe gratitude for various small and big favours, expert advice, and bioinformatic assistance. Last, but not least, I want to sincerely thank my former high school teacher Rainer Bürgel and my Diplom thesis advisor Christian Lehner for sparking and nourishing my interest in molecular biology.

I offer my gratitude to the Departments of Microbiology & Immunology and Earth & Ocean Sciences of the University of British Columbia as well as the Daimler-Benz Foundation, Germany for providing financial assistance.

To my wife Simone, without whom I would have never started this dissertation, and to my son Patrick, without whom I would have never finished it.

1 Introduction to Large Aquatic DNA Viruses

1.1 Viruses in the marine environment

The presence of viruses and virus-like particles (VLPs) in marine environments has been known for several decades. As early as 1971, Lee reported a viral infection in the freshwater red alga *Sirodotia tenuissima* [1]. In the late 1980s and early 1990s, the first reliable numbers about the abundance of aquatic viruses emerged [2-4], showing that there are typically ~$1x10^7$ viruses per ml of seawater (reviewed in [5, 6]). This makes viruses the most abundant biological entities in the ocean, outnumbering their hosts usually by an order of magnitude [7]. Viral abundance is correlated with primary productivity and usually decreases with distance from shore and increasing depth [7]. Viruses are major agents of mortality [8] and can terminate algal blooms, e.g. of the dimethyl sulphide producer *Emiliania huxleyi* [9, 10] or harmful algal bloom formers such as *Heterosigma akashiwo* [11, 12] and *Heterocapsa circularisquama* [13, 14]. Viruses play important roles in marine ecosystems by influencing the biodiversity of their hosts and driving global biogeochemical processes [15]. Moreover, the marine virosphere harbours an enormous genetic diversity and genes are frequently exchanged between different viruses as well as between viruses and their hosts [16-19]. Since bacteria constitute the numerically dominant group of cellular plankton, bacteriophages have been studied in great detail (reviewed in [20]). In contrast, protist-infecting viruses have only recently attracted wider attention, due to the fact that several of them contain very large double-stranded (ds) DNA genomes with hundreds of genes. Since many of these genes have homologues in cellular organisms, protistal viruses are prime candidates to study horizontal gene transfer, virus-host co-evolution, and the evolution of large DNA virus genomes.

1.2 Giant DNA viruses

Viruses are extremely diverse in terms of capsid size, capsid shape, type of nucleic acid and genome size [21]. The smallest virus genomes (*Circoviridae* with 1.8 kilobases [kb] single-stranded DNA) and the largest virus genomes (*Mimiviridae* with 1.2 megabase pair [Mbp] dsDNA) are separated by three orders of magnitude. For several decades, herpesviruses and poxviruses were the largest known viruses, until a family of double-stranded (ds) DNA viruses called the *Phycodnaviridae* was discovered in the 1980s [22]. These algal viruses have genome sizes of 155 kbp to 560 kbp and encode between 150 and 472 proteins [23]. Another relatively unknown family of large dsDNA viruses are the *Polydnaviridae*, which exist in a symbiotic relationship with parasitic wasps. They are normally excluded from comparative analysis with

other large dsDNA viruses owing to the proviral and multipartite state of their genomes, their extremely low gene densities, and their unique gene contents [24, 25]. The largest fully sequenced virus genome currently known is that of *Acanthamoeba polyphaga* mimivirus [26]. The term "giant virus" (also called girus [27]) generally refers to viruses with large dsDNA genomes, but is loosely defined with regard to the included genome range. Whereas some people set the lower threshold to ~150 kbp [28], others draw the line at 280-300 kbp [27, 29]. Alternatively, one could only consider viruses with genomes larger than the smallest cellular genomes, which would set the threshold to 500-600 kbp (*Nanoarchaeum equitans*: 491 kbp; *Mycoplasma genitalium:* 580 kbp). I will here use the term giant virus to refer to dsDNA viruses with a genome size larger than 300 kbp. According to this definition, there are currently ~24 giant viruses known (Table 1.1). This number may soon be doubled, as 19 new giant viruses have recently been isolated on *A. polyphaga* [30].

Table 1.1 Giant viruses.

Virus species [Reference]	Genome size (bp of dsDNA)	Predicted protein-coding genes	Coding density
Acanthamoeba castellanii mamavirus [31]	~1,200,000	N/A	N/A
Acanthamoeba polyphaga mimivirus [26]	1,181,404	981	90%
Cafeteria roenbergensis virus (this study) [32]	~730,000	544	90%
Glyptapanteles flavicoxis bracovirus [33]	~594,000 (in 29 segments)	193	32%
Cotesia congregata bracovirus [24]	~568,000 (in 30 segments)	156	27%
Pyramimonas orientalis virus [34]	~560,000	N/A	N/A
Glyptapanteles indiensis bracovirus [33]	~517,000 (in 29 segments)	197	33%
Chrysochromulina ericina virus [34]	~510,000	N/A	N/A
Bacillus megaterium phage G [35]	497,513	N/A	N/A
Phaeocystis pouchetii virus 01 [36]	~485,000	N/A	N/A
Phaeocystis globosa virus 1 [37]	~466,000	N/A	N/A
Emililania huxleyi virus 86 [10]	407,339	472	90%
Paramecium bursaria chlorella virus NY2A [38]	368,683	404	92%
Marseillevirus [39]	368,454	547	89%
Canarypox virus [40]	359,853	328	90%
Heterocapsa circularisquama virus [14]	~356,000	N/A	N/A
Paramecium bursaria chlorella virus AR158 [38]	344,690	360	92%
Ectocarpus fasciculatus virus a [41]	~340,000	N/A	N/A

Virus species [Reference]	Genome size (bp of dsDNA)	Predicted protein-coding genes	Coding density
Myriotrichia clavaeformis virus [42]	~340,000	N/A	N/A
Ectocarpus siliculosus virus 1 [43]	335,593	240	70%
Paramecium bursaria chlorella virus 1 [44]	330,743	366	90%
Paramecium bursaria chlorella virus FR483 [45]	321,240	335	93%
Pseudomonas chlororaphis phage 201φ2-1 [46]	316,674	462	93%
Paramecium bursaria chlorella virus MT325 [45]	314,335	331	N/A
Shrimp white spot syndrome virus 1 [47]	305,107	181	92%

1.2.1 The nucleocytoplasmic large DNA viruses

Although viruses as a whole are polyphyletic, there is evidence that at least some classes of viruses have evolved from a common ancestor. Examples of monophyletic groups are the positive-strand RNA viruses, the retroviruses and retroelements, the rolling-circle replicating viruses and mobile elements, and the tailed bacteriophages [48]. One of the most diverse groups of apparent monophyletic origin is the nucleocytoplasmic large DNA virus (NCDLV) clade. This clade consists of eukaryotic viruses with mostly icosahedral, but also bacilliform or pleomorphic capsids and dsDNA genomes larger than 100 kbp. The NCLDVs replicate partially or entirely in the host cytoplasm and comprise the families *Ascoviridae, Asfarviridae*, *Iridoviridae*, *Mimiviridae, Phycodnaviridae*, *Poxviridae*, and the Marseillevirus [39, 49, 50]. NCLDVs infect a wide range of host species, from amoebae and algae to insects and mammals. Table 1.2 lists selected characteristics of the individual NCLDV families.

Because NCLDVs comprise the largest known viruses and the number of fully sequenced NCLDV genomes is growing continually, this group of viruses provides a large dataset to study the evolution of large DNA viruses and the impact of viruses on the evolution of their hosts via horizontal gene transfer (HGT). Their classification into a presumably monophyletic group is based on a common set of genes that are present in most members and are referred to as NCLDV core genes. Since the first analysis by Iyer, Aravind, and Koonin [49], the presumed core gene set of the ancestral NCLDV has expanded from 31 to about 50 genes [51, 52]. The NCLDV ancestor must thus have been a virus of considerable size and complexity, encoding enzymes for genome replication, transcription, nucleotide metabolism, genome packaging, redox catalysis, and virion structure [49, 50, 52]. Some lineages have lost several of these ancestral genes, e.g. the phycodnavirus *Feldmannia sp.* virus 158, which carries a significantly reduced set of NCLDV core genes [28], and all phycodnaviruses, except for the

coccolithoviruses, which lack transcription genes [53]. Nevertheless, all NCLDVs have experienced a net genome growth over time and they comprise the most complex viruses known.

Mimivirus and the poxviruses replicate exclusively in the host cytoplasm and encode all genes necessary for DNA replication and transcription, as well as many additional enzymatic functions [54, 55]. Many NCLDVs possess genes that are usually absent in viruses, including enzymes involved in protein biosynthesis, DNA repair, and diverse metabolic pathways [50]. A detailed discussion of the extensive coding potential found in NCLDVs would go beyond the scope of this introduction, but several review articles have been published on this topic [49-52]. In a recent effort to delineate sets of NCLDV genes that may have evolved from a common ancestor, Yutin and coworkers constructed sets of orthologous genes for 11,468 predicted proteins, encoded by 45 NCLDV genomes [51]. The majority of these genes were assigned to 1,445 NCVOGs (Nucleocytoplasmic Virus Orthologous Genes) that were family-specific, which implies that there is little overlap in gene content among different NCLDV families. On one hand, this finding reflects the different evolutionary histories of individual virus families, owing to varying degrees of HGT from cellular sources. On the other hand, lineage-specific gene expansion resulted in large clusters of paralogous genes, e.g. the 66 ankyrin repeat-containing proteins in Mimivirus [56]. However, 177 NCVOGs contained genes from two or more NCLDV families and they represent the key functional classes of viral replication and virion assembly [51]. The five NCVOGS that were found in all analyzed NCLDV genomes are the major capsid protein, the DNA-dependent DNA polymerase B, the FtsK-HerA-like packaging ATPase, the archaeo-eukaryotic DNA primase-helicase, and the VV A2-like transcription factor.

Table 1.2 Family characteristics of nucleocytoplasmic large DNA viruses.

Family	Host range	Virion	Genome	CDSs	Virion proteins	Replication Site
Ascoviridae	insects	bullet-shaped, 130 nm x 200-240 nm	circular dsDNA, 130-186 kbp	123-180	N/A	nucleus + cytoplasm
Asfarviridae	mammals	icosahedral, 200 nm	linear dsDNA, 170-190 kbp	151	50	nucleus + cytoplasm
Iridoviridae	fish, amphibians, arthropods	icosahedral, 120-200 nm	linear dsDNA, 102-280 kbp	105-468	44	nucleus + cytoplasm
Marseillevirus, Lausannevirus[1]	*Acanthamoeba sp.*	icosahedral, 200-250 nm	circular dsDNA, 347-368 kbp	450-457	49	nucleus (?) + cytoplasm
Mimiviridae	*Acanthamoeba sp.*	icosahedral, 400-600 nm capsid with 125 nm outer fibre layer	linear dsDNA, 1.2-1.3 Mbp	~1000	137	cytoplasm
Phycodnaviridae	photosynthetic protists	icosahedral, 100-220 nm	linear/circular dsDNA, 155-560 kbp	150-472	28-125	nucleus + cytoplasm
Poxviridae	insects, birds, mammals, reptiles	pleomorphic, 140-300 nm	linear dsDNA, 134-360 kbp	133-328	80	cytoplasm

1.2.2 Mimivirus

Although the existence of giant viruses was known since the 1970s from electron microscopy of natural seawater samples [57], and despite two decades of research on phycodnaviruses [58], there was relatively little scientific and public interest in giant viruses. This changed radically with the discovery of *Acanthamoeba polyphaga* mimivirus. Isolated in 1992 from a cooling tower in England, Mimivirus was originally mistaken for a bacterium, due to its size and propensity to stain gram-positive [59]. Several years later, electron microscopy and partial genome sequencing revealed the true identity of Mimivirus as the largest virus seen to date [60]. The 1,181,404 bp dsDNA genome sequence was published in 2004 [26] and set off an international

[1] Lausannevirus was presented as a new strain of Marseillevirus at the 2010 "Viruses of Microbes" meeting at the Institut Pasteur, Paris, France. [100]

effort to study this behemoth of the viral world. As of October 2010, a search for “Mimivirus” retrieved 93 publications in PubMed[2].

Mimivirus has an icosahedral capsid with a diameter of 500 nm, which is covered by a 125 nm thick layer of fibres [61, 62]. The peripheral fibres may mimic a bacterial lipopolysaccharide layer in order to trigger phagocytic uptake by the amoeba [63]. The fibre layer also effectively increases the diameter of the virion, which may be necessary for phagocytosis, as the minimum diameter of individual particles to trigger phagocytosis is 0.5-0.6 µm [64, 65]. The capsid possesses a unique vertex with a starfish-like protrusion, also called the "stargate" structure [62, 66]. The stargate acts as a portal for genome delivery, whereby five icosahedral faces open up to release the virion core. Genome packaging proceeds through a temporary opening in an icosahedral face opposite of the stargate portal [66]. The virion contains two internal membranes [61]. After phagocytosis and stargate opening, the outer membrane fuses with the phagosome, releasing the genome into the cytoplasm, still enclosed within the inner membrane [66, 67]. Transcription of early genes then starts in this core vesicle [54], containing at least six RNA-dependent RNA polymerase subunits and additional transcription proteins [68]. The core eventually forms the cytoplasmic virion factory (VF), which produces progeny virions in an assembly line-like manner at its periphery, whereby the genome is injected into preformed capsids, followed by synthesis of the outer fibres [67]. Progeny virions are released from the cell through lysis, which usually occurs within 24 hours of infection.

The Mimivirus genome is a linear dsDNA molecule of 1.2 Mbp, which is predicted to adopt a Q-like secondary structure [59]. The initial genome analysis predicted 911 protein-coding sequences (CDSs) [26]; a subsequent transcriptomic study using mRNA deep sequencing identified further genes and increased the number of CDSs to 981 [69]. Roughly 30% of CDSs were assigned a putative function. An unexpected finding was the presence of a set of translation-related genes including 4 aminoacyl-tRNA synthetases, 5 translation factors, and 6 tRNA genes [26]. These genes were remarkable because viruses usually rely entirely on the host translation apparatus for protein synthesis. Several other Mimivirus CDSs had not been found in viruses before, including three DNA topoisomerases and the DNA repair enzymes formamidopyrimidine-DNA glycosylase, 6-O-methylguanine-DNA methyltransferase, and mismatch repair protein MutS. Furthermore, Mimivirus encodes a large DNA replication and transcription machinery, a partial glutamine metabolism pathway, six glycosyltransferases, heat shock proteins HSP40 and HSP70, and an intein element in the DNA polymerase B gene [26].

[2] http://www.ncbi.nlm.nih.gov/sites/entrez

Two promoter motifs were identified in the genome, one associated with early gene transcription [70] and one with late gene transcription [69]. The early gene promoter has a highly conserved consensus sequence (AAAATTGA) and seems to be specific for Mimivirus. Another peculiar finding about the mode of transcription in Mimivirus was that 85% of mature viral mRNAs had a characteristic hairpin structure at their 3' end [71]. The underlying intergenic palindromes, which may also cause transcription termination, are unique to Mimivirus. This combined evidence strongly indicates that Mimivirus gene transcription is entirely encoded by the viral genome and occurs in the cytoplasm, independent of nuclear enzymes.

Several Mimivirus gene products have been biochemically characterized, including tyrosyl and methionyl tRNA synthetases [72, 73], nucleoside diphosphate kinase [74, 75], cyclophilin [76], mRNA capping enzyme [77, 78], DNA ligase [79], DNA topoisomerase IB [80], endonuclease VIII [81] and a mitochondrial nucleotide carrier protein [82]. All of these studies confirmed that the viral proteins were enzymatically active, showing that despite its unusual size, the Mimivirus genome encodes genes that are functional and appear to evolve under selective pressure. Analysis of the Mimivirus virion proteome showed that this giant virus packages ~137 virally encoded proteins, including at least 15 transcription-related and 9 redox-active proteins [59, 68]. This makes Mimivirus a multiple record holder in the viral world, having the biggest virion, the largest viral genome, the most CDSs, and the most complex virion proteome.

Mimivirus is the prototype member of the *Mimiviridae* family. Based on phylogenetic reconstruction of a concatenated alignment from several universally conserved proteins, Mimivirus was proposed to represent a new domain of life, having descended from an extinct cellular lineage by genome reduction [26]. However, Mimivirus shares the conserved set of NCLDV core genes and is thus related to phycodnaviruses, poxviruses, and other NCLDV families, which makes it a bona fide virus. Because its genome and particle size were larger than the smallest unicellular organisms, Mimivirus has sparked an intense debate about the origin and evolution of DNA viruses and revived the interesting, albeit rather philosophical, question whether viruses should be considered alive [27, 83-92].

1.2.3 How common are giant viruses in the environment?

The discovery of Mimivirus raised the question whether giant viruses are widespread in natural environments and might thus play an ecological role. Observations of large VLPs in natural sea water samples date back to the 1970s. In 1979, Sicko-Goad and Walker described large VLPs in the dinoflagellate *Gymnodinium uberrimum* [57]. In this report, the VLPs measured 385 nm in diameter and had a pentagonal or hexagonal profile, suggesting icosahedral morphology.

Large VLPs ranging up to 750 nm in diameter have also been described in the vacuoles of some radiolarian groups isolated from environmental samples [93] as well as in Antarctic sea ice [94]. All of these viruses appeared to have icosahedral capsids with a diameter of 300 nm or more, containing an electron-dense core surrounded by a multilayer membrane system. Although none of the viruses have been isolated or further characterized, these reports provided a first hint at the existence of very large viruses in various marine and freshwater environments.

Taking advantage of the wealth of metagenomic data collected from various marine environments by Venter *et al.* [95, 96], 15% of Mimivirus genes were found to have their closest homologues in the Sargasso sea metagenomic dataset [97] and 25% of Mimivirus genes were significantly similar to sequences in Sorcerer II Expedition Global Ocean Sampling (GOS) metagenomic data set [98]. Focussing on the Mimivirus DNA polymerase B protein, 811 homologues from putative giant viruses were found in the GOS dataset [99]. However, these metagenomic datasets contain no information about the identity of the source organism, and the question of host range for giant viruses remained unanswered. Isolation efforts concentrated on *Acanthamoeba,* host for the largest known virus. It did not take long until a second Mimivirus strain was isolated from a cooling tower in Paris. This strain was named Mamavirus, had a slightly larger genome than Mimivirus, and was 99% similar in coding content to Mimivirus [31]. Another giant virus isolated on *Acanthamoeba* was Marseillevirus; it has a 368 kbp genome and constitutes the first member of a new NCLDV family [39]. Subsequently, a second strain named Lausannevirus was found to be very similar to Marseillevirus [100]. Recently, a screen for *Acanthamoeba*-infecting giant viruses in 105 samples from such diverse environments as seawater, rivers, cooling towers, soil, and even clinical material, resulted in 19 new virus isolates with particle sizes ranging from 150 nm to 600 nm [30]. Genome analysis of these viruses is underway and will significantly expand the current giant DNA virus dataset. *Acanthamoeba* is unlikely to be the sole host for giant viruses, as very large DNA viruses have been reported from diverse algae [34, 37] and heterotrophic protists [101]. It has become apparent, however, that phagotrophic protists are a suitable breeding ground for giant viruses, as they often contain eukaryotic, bacterial, and viral DNA within the same cell, which facilitates HGT [39, 102]. Giant viruses seem to be a diverse and abundant component of the virosphere, but due to sampling protocols that tend to exclude viruses larger than 0.2 μm [103], they have been overlooked until recently.

1.3 Virophages

1.3.1 The Sputnik virophage

In 2008, a new strain of Mimivirus was isolated from a cooling tower in Paris and called *Acanthamoeba castellanii* mamavirus [31]. Associated with Mamavirus was a novel, much smaller virus, which replicated within the Mamavirus virion factory and had a faster replication cycle than Mamavirus (particles appeared at 6 h and 8 h post infection, respectively). Co-infection of amoebae with Mamavirus and the smaller virus increased the survival rate of the host cells while causing defective Mamavirus capsid structures and a 70% decrease in infective Mamavirus particles [31]. Because the smaller virus did not replicate in the absence of Mamavirus, it qualified as a satellite virus and was named Sputnik (Russian for "satellite"). However, Sputnik was remarkably different from known satellite agents and La Scola *et al.* argued that, because of its unusual lifestyle as a true parasite of another virus, Sputnik represented a new class of satellite viruses. They therefore coined the term "virophage" for any virus “using the viral factory of another virus to propagate at the expense of its viral host” [31]. Analysis of the 18 kbp circular dsDNA genome of Sputnik revealed that some of its 21 protein-coding genes had homologues in archaea, bacteria, and eukaryotes. Most interestingly, three genes were apparently derived from its host virus as they had close relatives in Mimi- and Mamavirus. This suggested that virophages could be hitherto unrecognized vehicles of HGT between giant viruses and cellular organisms. The coding content of Sputnik included genes typically found in NCLDVs and bacteriophages such as a superfamily 3 helicase and an FtsK-HerA family DNA packaging ATPase. An integrase of the tyrosine recombinase family led La Scola *et al.* to propose the existence of integrated forms of virophages, so-called provirophages [31]. Although the major capsid protein (MCP) showed no sequence similarity to known capsid proteins, cryo-electron microscopy analysis of the Sputnik virion suggested that its MCP might adopt a double jelly-roll tertiary structure, similar to the MCPs of PBCV-1 and other viruses of the PRD1-adenovirus lineage [104]. The same study showed that the Sputnik particle consisted of an icosahedral capsid with a diameter of 74 nm and T=27 symmetry. Cryo-EM also confirmed that the virion had an internal lipid membrane, as suggested by the presence of the FtsK-HerA-like packaging ATPase.

The virophage concept posits that Sputnik and similar agents are parasites of their host virus rather than of the host cell. This leads to the prediction that virophages use the DNA replication and transcription enzymes encoded by their host virus and present in the virion factory [105]. Support for this hypothesis came from analysis of the Mimivirus transcriptome, which identified

a putative late gene promoter motif in Mimivirus that was also found upstream of 12 Sputnik genes, suggesting that Sputnik genes are transcribed concomitantly with Mimivirus late genes [69]. Furthermore, transcriptional termination in Sputnik is probably governed by the Mimivirus transcription machinery. Eighty-five percent of Mimivirus transcripts were found to end in characteristic hairpin structures that were encoded by palindromic sequences [71]. This putative transcription termination or polyadenylation signal seems to be unique to Mimivirus – with the exception of Sputnik. Within its intergenic regions, Sputnik contains the same type of palindromes [105]. The finding that Sputnik encodes Mimivirus-specific transcription signals suggested that Sputnik might indeed be dependent on Mimivirus enzymes for gene expression.

1.3.2 Virophages and other satellite viruses

According to the International Committee of Taxonomy of Viruses (ICTV), "satellites are sub-viral agents composed of nucleic acid molecules that depend for their productive multiplication on co-infection of a host cell with a helper virus" [21]. Satellites encoding their own coat protein are called satellite viruses. The vast majority of satellite viruses occur in plants and have single-stranded RNA genomes, such as Tobacco mosaic satellite virus or chronic bee-paralysis satellite virus[3]. RNA satellite viruses are completely unrelated to the Sputnik virophage.

Of several satellite bacteriophages that have been discovered [106-108], the best known example is enterobacteria satellite virus P4. Its helper phage P2 provides proteins for P4 head and tail structure, genome packaging, and host cell lysis [109]. Genome replication of P4, on the other hand, occurs autonomously. The P4 genome is an 11,624 bp linear dsDNA molecule that circularizes after injection into the host cell. P4 encodes an integrase and, in the absence of a helper phage, can persist in the host as a prophage, or episomally as a high copy number plasmid [110]. In P4-lysogenic cells, the satellite phage can sense the presence of the helper and reactivate upon P2 superinfection. Because of its episomal state, it has been proposed that P4 may have evolved from a plasmid [110].

The only satellite viruses that might be remotely comparable to the Sputnik virophage are adeno-associated viruses (AAVs) of the dependovirus genus (family *Parvoviridae*). AAVs consist of a linear 4.7 kb ssDNA genome wrapped in a 22 nm diameter, non-enveloped icosahedral capsid. The AAV genome carries two genes (rep and cap) which can produce seven different proteins through the use of overlapping reading frames, alternative promoters, alternative splicing, and non-canonical start codons. The rep proteins are regulators of DNA replication and transcription, while the cap proteins constitute the capsid structure. AAVs are

[3] http://www.ictvdb.org/Ictv/fs_satel.htm

able to infect a wide range of mammalian cells and they may enter their host cells by receptor-mediated endocytosis via the formation of clathrin-coated pits. Dependoviruses have lytic or lysogenic infection cycles. Lytic infection depends on the presence of a helper virus such as adenovirus, herpes simplex virus, or cytomegalovirus ([111] and references therein), whereas in the absence of a helper virus, AAV can integrate its genome into the host chromosome, where it is passed on to subsequent generations [112]. Despite having small ssDNA genomes, the AAVs share several features with virophages. First, they are replication-dependent on medium-sized to large dsDNA viruses. Second, they negatively affect replication of their helper adenovirus in co-infected cells [111, 113]. Third, AAV can exist as a provirus, which has been proposed for virophages, but has not been demonstrated yet [31]. However, AAV genome integration is a passive process and AAVs do not encode an integrase [114]. AAVs and the Sputnik virophage have completely different genomes, and the presumably biggest difference is that AAV and its helper adenovirus replicate in the host nucleus, whereas Sputnik and Mimivirus replicate in cytoplasmic virion factories. Moreover, AAV promoters are not similar to their helper virus promoters and genes encoded by AAVs are not similar to adenoviral genes, as opposed to the observed Sputnik-Mimivirus promoter homology [69] and three Sputnik genes with close homologues in Mimivirus [31].

It can thus be concluded that Sputnik and potentially other virophages yet to be discovered are indeed very different from known satellite viruses and constitute a novel class of satellite agents.

1.4 Origin and evolution of viruses

Viruses appear to infect every organism on earth, from the smallest bacteria to the largest mammals [21]. This perception suggests that viruses must have played an enormous role in the evolution of their hosts, by direct genetic interaction as well as by imposing selective pressure on their host cells and controlling population growth. Yet, the origin and evolution of viruses themselves represents a methodologically bigger challenge than the origin and evolution of cellular life. First, unlike cellular organism, viruses leave no fossil records which could be used for dating ancestral forms. Second, whereas bacteria, mammals, and all other cellular life forms have descended from a last universal common ancestor (LUCA) [115, 116] and can be arranged in a tree-like fashion to represent their evolutionary descent, this is not possible for viruses. There is no single gene common to all viruses, which makes it impossible to reconstruct a phylogeny that would include all extant viruses. Instead, evolutionary virologists must restrict their analyses to subsets of viruses and use phylogenetic marker genes that are specific for these groups, e.g. the RNA-dependent RNA polymerase for positive-sense single-

stranded RNA viruses or DNA-dependent DNA polymerase B for NCLDVs. The extreme diversity of viruses, consisting of single- or double-stranded RNA or DNA genomes with various topologies, suggests that viruses are polyphyletic, i.e. they likely originated not just once, but multiple times.

Hypotheses to explain the origin of viruses can be classified into three categories: (i) The reduction hypothesis states that viruses were once cellular organisms which became intracellular parasites and consequently lost all genes for energy metabolism and protein translation [26, 117, 118]. The major caveat of this hypothesis is the existence of viral genes that have no cellular homologues, including widely distributed ones such as capsid genes. Hence, the reduction hypothesis proposes that the cellular lineage from which viruses descended has gone extinct. (ii) The escape hypothesis can be viewed as the "traditional" model of viral evolution and assumes that viral genomes are derived from cellular genome fragments that became autonomous and infectious [83, 117, 119]. However, comparative viral genomics and the unresolved issue of the origin of the capsid protein argue against this scenario [48]. The escape model is still vividly defended by some researchers who argue, for instance, that genomes of giant DNA viruses such as Mimivirus evolved by extensive gene acquisition from cells [84, 90]. (iii) The virus-first hypothesis suggests that viruses originated in a pre-cellular environment, and that viral genes that lack cellular homologues have originated in viruses [120, 121]. This hypothesis differs from the first two scenarios in the existence of "true" viral genes, which have never been part of a cellular genome. Some of these exclusively viral genes are well conserved and found in a wide variety of viruses. They are called viral hallmark genes (VHGs) and include the jelly-roll capsid (JRC) protein, the superfamily 3 helicase, and at least six additional proteins involved in genome replication and packaging [48]. The JRC protein is an especially interesting case, as it is found in viruses infecting archaea, bacteria, and eukaryotes, which implies that viruses encoding the JRC gene are at least as old as LUCA and could even have descended from a single common ancestor [122, 123]. On one hand, the capsid protein (there are at least three major versions of it, the double-jelly roll or PRD1 fold, the HK97 fold, and the blue tongue virus fold [124]) is the single unifying criterion that sets viruses apart from cells. Thus, Raoult and Forterre [85] proposed that the living world can be divided into ribosome-encoding organisms (REOs, i.e. cells) and capsid-encoding organisms (CEOs, i.e. viruses). On the other hand, viruses are obligate intracellular parasites and their evolution is tightly linked to the evolution of their hosts. Any hypothesis about the origin of viruses should thus be viewed in the light of cellular evolution, and vice versa. A hypothesis of an ancient viral lineage that is based solely on structural comparison of a single protein such as the JRC protein, however, ignores the crucial context of cellular evolution and fails to explain the more

complex distribution patterns of other VHGs [48]. According to Koonin *et al*., the existence of VHGs can be explained by an ancient virus world, in which the progenitors of modern cells were parasitized by primitive selfish genetic elements, which exchanged genes frequently and eventually invented the capsid protein to facilitate their horizontal transmission [48]. The nature of these parasitic agents changed with the evolutionary stage of their hosts, starting with RNA elements, and progressing via reverse-transcribing RNA/DNA elements to small DNA entities replicating through a rolling-circle mechanism, and finally to larger dsDNA viruses. This scenario implies that even the earliest stages of cellular evolution were tightly coupled with viral evolution. Because the currently predominant theory for the origin of eukaryotes posits a fusion between an archaeon and an α-proteobacterium [125], the genomes of eukaryotic viruses are predicted to have arisen from a combination of mostly bacteriophage and some archaeal virus genes [48].

During the past few years, discoveries of giant viruses encoding translation proteins and other typically cellular genes [10, 26], combined with metagenomic studies revealing the vast abundance of viruses in the ocean [126] and new insights about beneficial roles of retroviral genes in our own genomes [127, 128], have caused a shift in the public perception about viruses, from mere agents of disease to active players in evolution and ecology. As an example, it has been pointed out by several authors that viruses should not be confused with virions (i.e. the virus particle), and that the intracellular virion factory (VF) represents the main phase of the viral life cycle (the viral "self") [118, 129-131]. The debate about the nature of viruses culminated in the virocell concept, in which Forterre compares the VF to an intracellular parasite and views the infected cell as a living, membrane-bound, viral organism, called the virocell [131].

To summarize, multiple hypotheses have been formulated to explain the origin and evolution of viruses, but to date, none of them can provide a satisfactory answer to this complex problem. However, with a steadily increasing number of fully sequenced viral genomes, comparative viral genomics has the potential to shed more light on the fascinating topic of viral evolution.

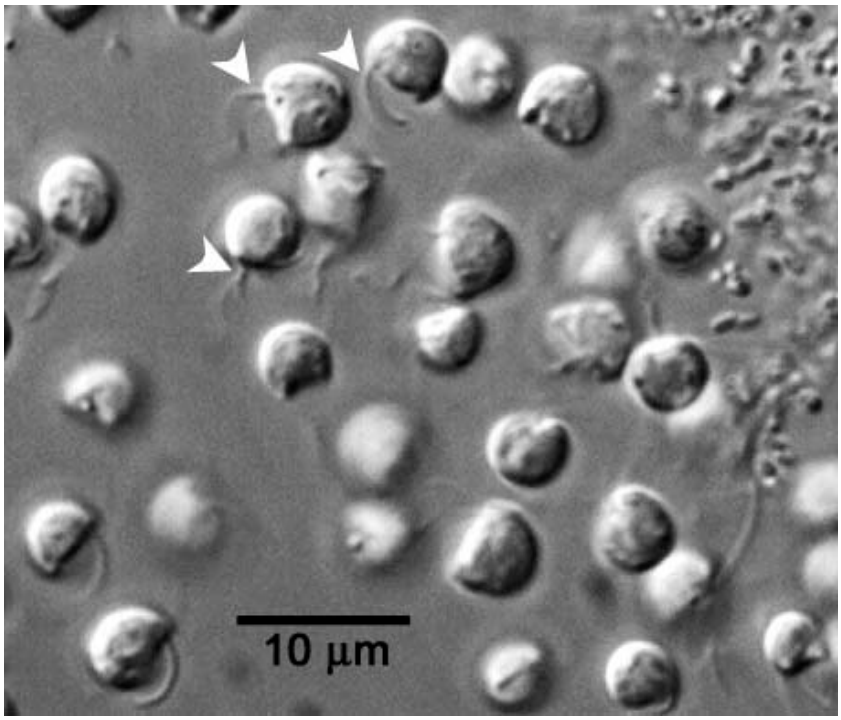
1.5 *Cafeteria roenbergensis* virus

1.5.1 *Cafeteria* – a genus of heterotrophic nanoflagellates

Heterotrophic flagellates are an integral part of the marine food web and dominate among protozoan grazers [132, 133]. They feed mainly on bacteria but can also ingest small phytoplankton, cell fragments, larger viruses, and particulate organic matter [134, 135].

One of the most common and abundant heterotrophic nanoflagellates in marine environments is the genus *Cafeteria* with its type species *C. roenbergensis*, first described by Fenchel and Patterson in 1988 [136]. *C. roenbergensis* has been found in various habitats such as surface waters, deep sea sediments, and hydrothermal vents [137, 138]. These 3-6 µm long, ellipsoid-shaped protozoa possess two flagella; the anterior flagellum is used for forward-propelling motion, the posterior flagellum is used for attachment to substrates. *C. roenbergensis* is an interception feeder, i.e. it encounters its prey by active swimming and engulfs it via membrane invagination [139, 140]. *Cafeteria* cells reproduce asexually by binary fission. Higher taxonomic classification places *Cafeteria* within the order Bicosoecales as part of the heterokonts (also referred to as Stramenopiles), a diverse group of eukaryotes with phototrophic, heterotrophic, and mixotrophic members. According to Cavalier-Smith [141], the Heterokonta belong to the kingdom Chromista, whose common ancestor acquired a red algal plastid that was consequently lost in several lineages. The protistan host for *Cafeteria roenbergensis* virus (CroV) was originally thought to belong to the genus *Bodo* [101], but partial 18S rDNA sequencing revealed 100% sequence identity to *C. roenbergensis* strain VENT1 [138]. A light microscopy image of *C. roenbergensis* strain E4-10 is shown in Fig. 1.1.

Figure 1.1 The heterotrophic nanoflagellate *Cafeteria roenbergensis* strain E4-10.
This light micrograph shows several cells of *C. roenbergensis* next to a patch of bacteria (upper right). Note the two flagella emerging from the flattened side of the cell (arrowheads). Flagellates were concentrated by centrifugation prior to microscopy. Image courtesy of Dr. Naoji Yubuki, University of British Columbia.



1.5.2 Viruses infecting *Cafeteria roenbergensis*

Predation by protistan grazers is a major pathway of carbon transfer and nutrient recycling in marine and freshwater systems [132]; yet, viruses infecting phagotrophic protists in marine systems are largely unknown and completely unexplored genetically. The large dsDNA virus that became known as CroV was originally isolated from Texas coastal waters in 1990 by M.Sc. student Randy Garza. In 1995, Garza and Suttle reported the discovery of a large lytic virus that infected the heterotrophic nanoflagellate strains E1 and E4 [101], which had been isolated from the Oregon coast [135]. They named this virus BV-PW1, for *Bodo* Virus Pier Water isolate 1. The icosahedral capsid had a diameter of 230-300 nm and contained dsDNA. The double-layered capsid contained an electron-dense core, and the presence of a lipid component was suggested by sensitivity of the virions to chloroform [101]. Virus particles accumulated in the cytoplasm of infected cells, with no evidence of nuclear assembly. Based on these observations, BV-PW1 was thought be a member of the NCLDV clade. A few years later, M.Sc. student Tanya St. John and Dr. Andrew Lang in the laboratory of Curtis Suttle developed a purification protocol for the virus and reported that the virion consisted of two major protein components with masses of 55 kDa and 65 kDa, and at least 30 minor proteins [142]. Pulsed-field gel electrophoresis of viral DNA suggested a genome size larger than 100 kbp and partial sequence analysis of a helicase gene amplified by random PCR hinted at African swine fever virus as the closest relative [142].

Although nothing is known about the ecology or biogeography of *Cafeteria*-infecting viruses, the finding that a virus isolated from the Gulf of Mexico was able to lyse a *C. roenbergensis* strain from the Oregon coast suggests that these viruses may be as widespread as their hosts. Furthermore, in a recent report, Massana *et al*. described the viral infection of a *C. roenbergensis* enrichment culture from the central Indian Ocean by large VLPs [143]. This is remarkable because the sample site was located in an oligotrophic region of the Indian Ocean with very little biomass available. Addition of low amounts of organic matter to 3 µm filtered surface seawater incubations stimulated the growth of bacteria and heterotrophic flagellates, and at day 4 of the incubation, *Cafeteria roenbergensis* dominated the cultures with up to 30,000 cells per ml. At day 5, 72% of the flagellates appeared infected by large (~280 nm diameter) VLPs and at day 7, the induced *C. roenbergensis* bloom had been terminated by the virus [143]. This demonstrates that viruses infecting *C. roenbergensis* are not only present in the Gulf of Mexico, but also occur in the Indian Ocean, which implies that populations of this heterotrophic nanoflagellate may be controlled by viruses worldwide.

1.6 Thesis objectives

Based on preliminary results about CroV, which suggested that the virus might be a new representative of the NCLDV clade, further characterization promised to yield interesting results about CroV and viruses infecting marine protists. The overall goal of this dissertation was to present a detailed characterization of the novel marine virus CroV on different biological levels: the infection cycle, the genome, the transcriptome, and the virion proteome. During a late stage of the project, the discovery of the Mavirus virophage revealed an additional key aspect of CroV biology.

This dissertation had the following objectives:

1. Analysis of the CroV genome
2. Analysis of the CroV transcriptome
3. Analysis of the protein composition of the CroV particle
4. Description of the CroV infection cycle
5. Characterization of the CroV-associated virophage Mavirus

To address these goals, the dissertation is structured as follows:

Chapter 2 analyzes the 730 kbp genome sequence of CroV, explores the phylogenetic relationship of selected genes of interest, and makes use of microarray data to categorize CroV genes into early and late genes.

Chapter 3 describes the protein composition of the CroV virion as determined by liquid chromatography tandem mass spectrometry, and compares it to virion proteomes of other giant viruses.

Chapter 4 describes the genome of the Mavirus virophage, a novel satellite virus that parasitizes CroV. This chapter also presents evidence for the viral transposogenesis hypothesis, stating that the *Maverick/Polinton* class of eukaryotic DNA transposons has originated from endogenization of ancient virophages related to Mavirus.

Chapter 5 describes ultrastructural aspects of the infection cycles of CroV and Mavirus by using thin-section transmission electron microscopy.

Chapter 6 integrates the results from genome, transcriptome, proteome, and ultrastructural analyses into a coherent picture about the biology of CroV and Mavirus and discusses some major implications arising from this work.

2 Genome Analysis of CroV[4]

2.1 Synopsis

This chapter describes the genome features of CroV, the first virus characterized to infect a marine zooplankton species. With ~730 kilobase pairs (kbp) of double-stranded (ds) DNA, CroV has the largest genome of any described marine virus and the second largest viral genome overall. The central 618 kbp 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 genes, respectively. Microarray analysis showed that 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 have not been found before in a virus. A 38 kbp genomic region of putative bacterial origin was discovered, which encodes 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 (NCLDV) and has *Acanthamoeba polyphaga* mimivirus as its closest relative, although less than a third of its genes have homologues in Mimivirus.

2.2 Materials and methods

2.2.1 Flagellate growth and virus purification

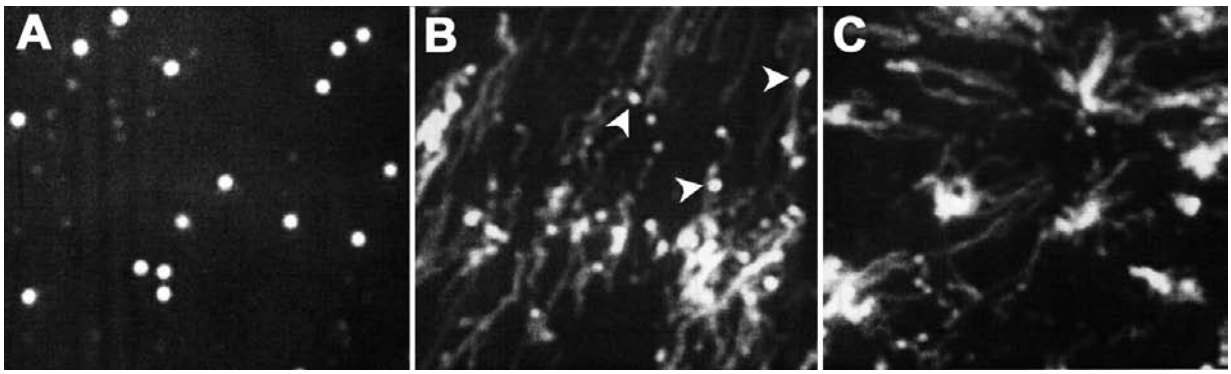
Cafeteria roenbergensis strain E4-10 was isolated from coastal waters near Yaquina Bay, OR, USA, as described previously [135]. Cultures of *C. roenbergensis* were grown in modified f/2 enriched seawater medium which consisted of 8.25×10^{-6} M Na_2EDTA, 8.25×10^{-6} M $FeCl_3$-EDTA, 5.00×10^{-6} M K_2HPO_4, 5.50×10^{-4} M $NaNO_3$, 1.00×10^{-8} M Na_2SeO_3, 3.93×10^{-8} M $CuSO_4$, 5.87×10^{-9} M $CuCl_2$, 2.60×10^{-8} M Na_2MoO_4, 7.65×10^{-8} M $ZnSO_4$, 9.10×10^{-7} M $MnCl_2$, 2.96×10^{-7} M thiamine, 2.05×10^{-9} M biotin, and 3.69×10^{-10} M cyanocobalamin in 30 kDa MWCO ultrafiltrated seawater of ~30 ppt salinity. Following autoclaving, the f/2 medium was supplemented with 0.01% (w/v) yeast extract (Difco BD, Sparks, MD, USA) to stimulate the growth of a mixed bacterial assembly that was present in the cultures.These marine bacteria grew to high concentrations of

[4] A version of this chapter has been published: Fischer M.G., Allen M.J., Wilson W.H. and Suttle C.A. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 107:19508-19513.

up to $5x10^8$ cells/ml and served as the primary food source for *C. roenbergensis.* Cultures were kept at room temperature (~22°C) in the dark. Typically, 250 ml of exponentially growing *C. roenbergensis* in one-litre plastic Erlenmeyer flasks were infected at a cell density of $5x10^4$ cells/ml by adding 100 µl (~0.01 multiplicity of infection [MOI]) of CroV-containing crude lysate. Flagellate growth was monitored by staining cells with Lugol's Acid Iodine and counting by light microscopy on a hemocytometer, which had a detection limit of $1x10^3$ cells/ml. Lysates were centrifuged for 1 h 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 h 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% (w/v in 50 mM Tris-HCl, pH 7.6) sucrose gradient, and centrifuged for 1 h 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, diluted 1:1 with sterile 50 mM Tris-HCl, pH 7.6, and centrifuged for 1 h at 150,000 x g (SW40 rotor, 20˚C). Virion pellets were resuspended in sterile 50 mM Tris-HCl, pH 7.6 and stored at 4˚C. Glutaraldehyde-fixed (0.5% w/v) virus was quantified by epifluorescence microscopy (SYBR Green I, Invitrogen; Whatman Anodisc filter membranes, VWR Canada) as described [144].

Figure 2.1 Denaturation of CroV particles.

(**A**) Intact CroV virions before treatment. (**B**) CroV virions after 20 hours at 55°C in the presence of 1% (w/v) lauroylsarcosine. Long strands of free viral genomic DNA are intermingled with remaining intact particles (arrowheads). (**C**) CroV virions after 20 hours at 55°C in the presence of 1% (w/v) lauroylsarcosine and 1 mg/ml Proteinase K. The combination of heat, detergent, and Proteinase K resulted in complete denaturation of the virions. All treatments were performed in 20 µl reactions which were then filtered onto 0.02 µm pore size, 25 mm diameter Anodisc Al_2O_3 filters and subsequently stained with SYBR Green I and examined at 500-fold magnification under an epifluorescence microscope.



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% (w/v) N-lauroylsarcosine and 1 mg/ml Proteinase K and incubated at 55˚C for at least 12 h. Treatment by heat and detergent in the absence of Proteinase K resulted in only partial virion denaturation, as shown in Fig. 2.1.

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 100% ethanol. After centrifugation, the DNA pellet was dissolved in nuclease-free molecular grade water (Invitrogen, Burlington, ON, Canada). Quantification and quality control were performed on a Nanodrop spectrophotometer.
Alternative virus purification and DNA extraction methods are described in section 3.2.1.

2.2.2 Pulsed-field gel electrophoresis of CroV genomic DNA

Chromosome length was analyzed by pulsed-field gel electrophoresis (PFGE). Approximately $5x10^8$ virions (purified by density gradient centrifugation) were embedded in 1% (w/v) 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 supplemented with 1% (w/v) lauroylsarcosine and 1 mg/ml Proteinase K, at 50°C overnight to release the viral genome from the capsid. The gel plug was then washed three times with 1x TE, pH 7.6 for 30 minutes each and once with 0.5x TBE and sealed in a 1% (w/v) agarose gel in 0.5x TBE. PFGE run conditions were 25 hours at 200 V (6 V/cm) and 14°C, with 60-120 sec switch times at a 120° angle and a ramping factor of 16.17 using a BioRad DR2 CHEF unit. DNA bands were visualized by ethidium bromide staining. 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 digest, ~$5x10^9$ 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 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 with 100 µg/ml 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 h and then for another 8 h in a fresh reaction mixture. Following digestion, gel plugs were incubated for 2 h at 50°C in 300 µl of L buffer with 1 mg/ml Proteinase K. Gel plugs were rinsed three times with 1x TBE and sealed in a 1% agarose gel. PFGE conditions were as described above, with run and switch times varying according to DNA fragment lengths.

2.2.3 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 megabases 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 GS 20 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 megabases 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 coverage of 38-fold. The 38 large GS FLX contigs that were associated with the CroV genome comprised 601,547 bp of non-redundant sequence, with an average contig size of 15,797 bp.

To span the inter-contig regions, multiple oligonucleotide primers were designed for each contig end, their 3' ends distally oriented. Different primer combinations were used in multiplex polymerase chain reactions (PCR) containing up to 12 different primers. Reactions yielding one or more distinct bands were repeated with smaller subsets of the original primer mix to identify which primers were responsible for the PCR product. Once identified, 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. PCR assays were conducted in 25 µl or 50 µl reactions, containing 1-2 pg/µl CroV genomic DNA, 1 µM of each primer, 1.5 µM $MgCl_2$ (up to 4.5 µM in multiplex reactions), 0.2 mM dNTPs, 0.02 U/µl Platinum Taq DNA polymerase (Invitrogen), and 1x Platinum Taq reaction buffer. Thermocycling protocols consisted of 5 min at 95°C (1 cycle); 30 sec at 95°C, 1 min at a suitable annealing temperature (usually 53-56°C), 3 min at 72°C with a maximum heating rate of 1°C/sec (35 cycles); 12 min at 72°C (1 cycle), and cooling at 4°C. A list of primer sequences is available in Appendix D. Several regions of the final genome assembly, mainly those containing DNA repeats as well as all contig breakpoints, were re-sequenced to increase coverage. A small insert shotgun library was created to aid in genome sequencing and turned

out be valuable for 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% (w/v) 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 kbp 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 kbp insert size], CopyControl Fosmid Library Production Kit, Epicentre, Madison, WI, USA; and pJAZZ-KA vector [10-20 kbp insert size], Lucigen) was unsuccessful despite positive tests with provided insert DNA and the successful cloning of phage307 genome fragments (phage307 is described in Appendix E).

2.2.4 Genome annotation

Artemis software v12.0 [145] was used for genome annotation. CDSs predicted and numbered by Artemis were compared to those predicted by the EMBOSS application GETORF [146]. A CDS was defined 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 which was only 47 codons long, see section 3.3.1). 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 methionine. 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 [147] with an E-value cutoff of 10^{-5}, and the COG database [148] 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 [149] and InterPro [150] databases. In cases where these predictions were still ambiguous or inconclusive, multiple sequence alignments with putative homologues 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*.[5] [51]. Hits with E-values < 10^{-5} were assigned to their respective NCVOGs. Putative tRNA genes were identified with tRNAscan-SE using the general

[5] downloaded from ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG

tRNA model [151]; codon analysis was carried out using CodonW[6]. Inteins were identified in BLAST searches based on gaps in alignments with intein-free homologues. Detailed sequence alignments with allelic inteins and identification of conserved amino acids corroborated the annotation as theoretical inteins. 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 [153]. 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, I 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, I 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 [154].

2.2.5 Phylogenetic analysis

For phylogenetic reconstruction, the taxonomic distribution of the query protein was examined using the InterPro database [150]. Putative homologues were identified by separate BLASTP searches against viruses, eukaryotes, bacteria, and archaea, or, where necessary, taxonomic subgroups thereof. Upon visual inspection of the potential homologues, 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 [155], followed by manual refinement. Maximum likelihood (ML) trees were generated on the phylogeny.fr webserver [156] or the PHYML webserver [157]. Bayesian Inference (BI) analysis as implemented in MrBayes v3.1.2 [158] was carried out with gamma rates and the mixed amino acid model. MrBayes was run for at least 1 million generations or until the standard deviation of split frequencies was less than 0.01. BI

[6] http://mobyle.pasteur.fr/cgi-bin/MobylePortal/ portal.py?form=codonw

trees were generated by the majority rule consensus method, which produces an output tree that contains only nodes that are present in at least half of the input trees.

2.2.6 Microarray analysis[7]

To create the microarray, PICKY software [159] was used to design one 50-70 bp long oligonucleotide each for 438 of the 544 predicted CroV CDSs. The remaining CDSs had either not yet been annotated or were too repetitive for unique probe design. 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 ("spot aliens" and "landing lights").

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 (immediately after infection), 1, 2, 3, 6, 12, 24, 48, and 72 h p.i.. It should be noted that although the CroV infection cycle lasted 12-18 hours, cultures frequently contained living cells up to 5 days post infection. This was due to the low MOI used, which required 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

[7] Microarray analysis was carried under supervision of Dr. Michael J. Allen in the laboratory of Dr. William H. Wilson at the Plymouth Marine Laboratories, United Kingdom.

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 in a final volume of 0.5 ml. 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 were 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 double-stranded 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.

Labelling of cDNA with fluorescent dyes

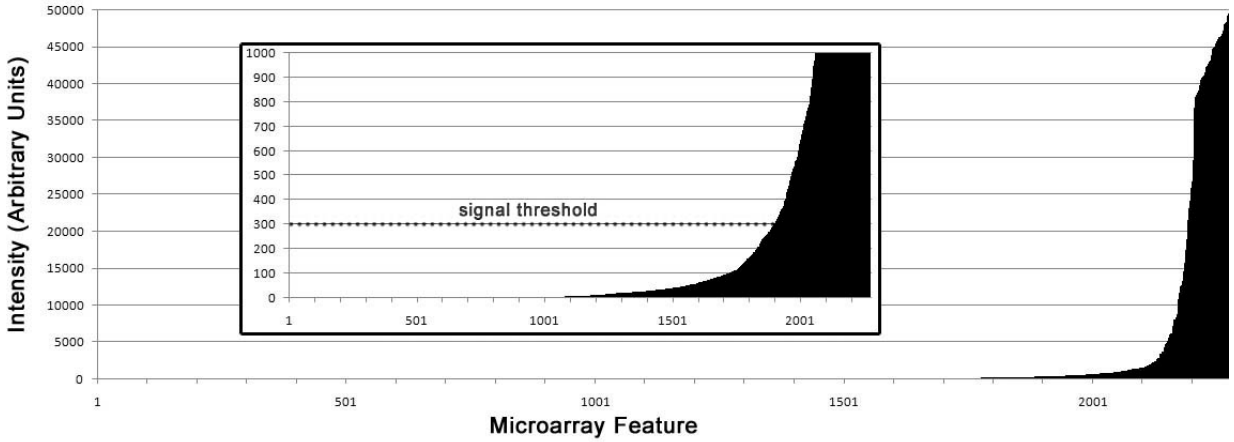
PCR-amplified cDNA was labelled with Cy3 by *in vitro* transcription using the Microarray Target Synthesis Kit (Roche, UK). 10.75 µl of template DNA were combined with 2 µl DTT (100 mM), 1 µl NTP mix (25 mM ATP, 25 mM CTP, 25 mM GTP, 18.75 mM UTP), 1.25 µl Cy3-17-UTP (5 mM, Amersham, UK), 2 µl transcription buffer (10x) and 3 µl transcription enzyme blend. The reaction was incubated for 16 hours at 37ºC and Cy3-labelled 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 labelled 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 labelled 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 [160] 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. 2.2. 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 labelled cRNA isolated from the uninfected control culture. These values were chosen based on experience from previous microarray studies [160]. 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 were collected to be MIAME compliant.

Figure 2.2 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.



Host strain18S Sequencing

Eukaryotic 18S rDNA fragments were amplified from *C. roenbergensis* using universal eukaryotic primers Euk1A and Euk516r as previously described [161]. 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.

Data deposition

The sequences reported in this chapter have been deposited in the GenBank database [accession numbers 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[8] under the accession number GSE19051. The four CroV intein sequences have been submitted to InBase, the NEB intein database[9].

[8] http:// www.ncbi.nlm.nih.gov/geo
[9] http://www.neb.com/neb/inteins.html

2.3 Results and discussion

2.3.1 General genome features

The CroV genome is a dsDNA molecule with a size of ~730 kbp, as determined by PFGE analysis (Figs. 2.3, 2.4), making this the second largest sequenced viral genome after *Acanthamoeba polyphaga* mimivirus. The CroV genome was sequenced by 454 pyro-sequencing and assembled *de novo* using GS assembly software in combination with other methods (see section 2.2.3 and Appendix A for details). The viral chromosome consists of a central 618 kbp protein-coding part, which is flanked on both ends by extensive regions of putatively repetitive nature (Fig. 2.5). 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 (76.6% 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 about 10% of the overall aa (Fig. 2.6).

Whole-genome restriction analysis was carried out to determine whether the genome had linear or circular topology (Fig. 2.4). Purified virus particles were embedded in agarose blocks and the viral genome was released by protease/detergent/heat treatment (Fig. 2.1). In-gel digestion was then performed separately using three different restriction endonucleases: *Apa*I was predicted to cut the genome at two positions, and *Fsp*I and *Sac*II were predicted to cut the genome once each (Table 2.1). Restriction digests were analyzed by PFGE, as shown in Fig. 2.4. Due to the wide size range of predicted DNA fragments, the gel was run in three consecutive steps, with photographs taken after each completed step. Whereas *Apa*I digested the genome quantitatively (note the absence of a full-length genomic band in Fig. 2.4C), *Fsp*I had a cut efficiency of less than 100% (some undigested DNA visible, band 4 in Fig. 2.4C) and *Sac*II had very poor cut efficiency (most of the DNA remained undigested, band 10 in Fig. 2.4C). These partial restriction digests may be caused by methylation of the CroV genome, as the virus encodes several predicted methyltransferases (see section 2.3.5). Due to prolonged times of electrophoresis, staining, and destaining, as well as the large amount of DNA used to ensure good band visibility, the size estimates resulting from this restriction digest were afflicted with large error margins. Therefore, the pulsed-field gel shown in Fig. 2.3, which had a much better resolution, was used to estimate the genome size at 730 kbp.

Nevertheless, all three restriction digests shown in Fig. 2.4 confirmed the linearity of the CroV genome, as two restriction fragments were observed for enzymes that cut the genome once, and three restriction fragments were observed for *Apa*I which had two predicted target sites in the CroV genome (Table 2.1).

Figure 2.3 Pulsed-field gel electrophoresis of CroV genomic DNA.
DNA from approximately 1x10^9 virions was separated on a 1% agarose pulsed-field gel. *Saccharomyces cerevisiae* chromosomes and lambda DNA concatemers were used as DNA size markers. Electrophoresis was carried out at for 25 h at 6 V/cm, 14°C, and 60-120 sec ramped switch times.

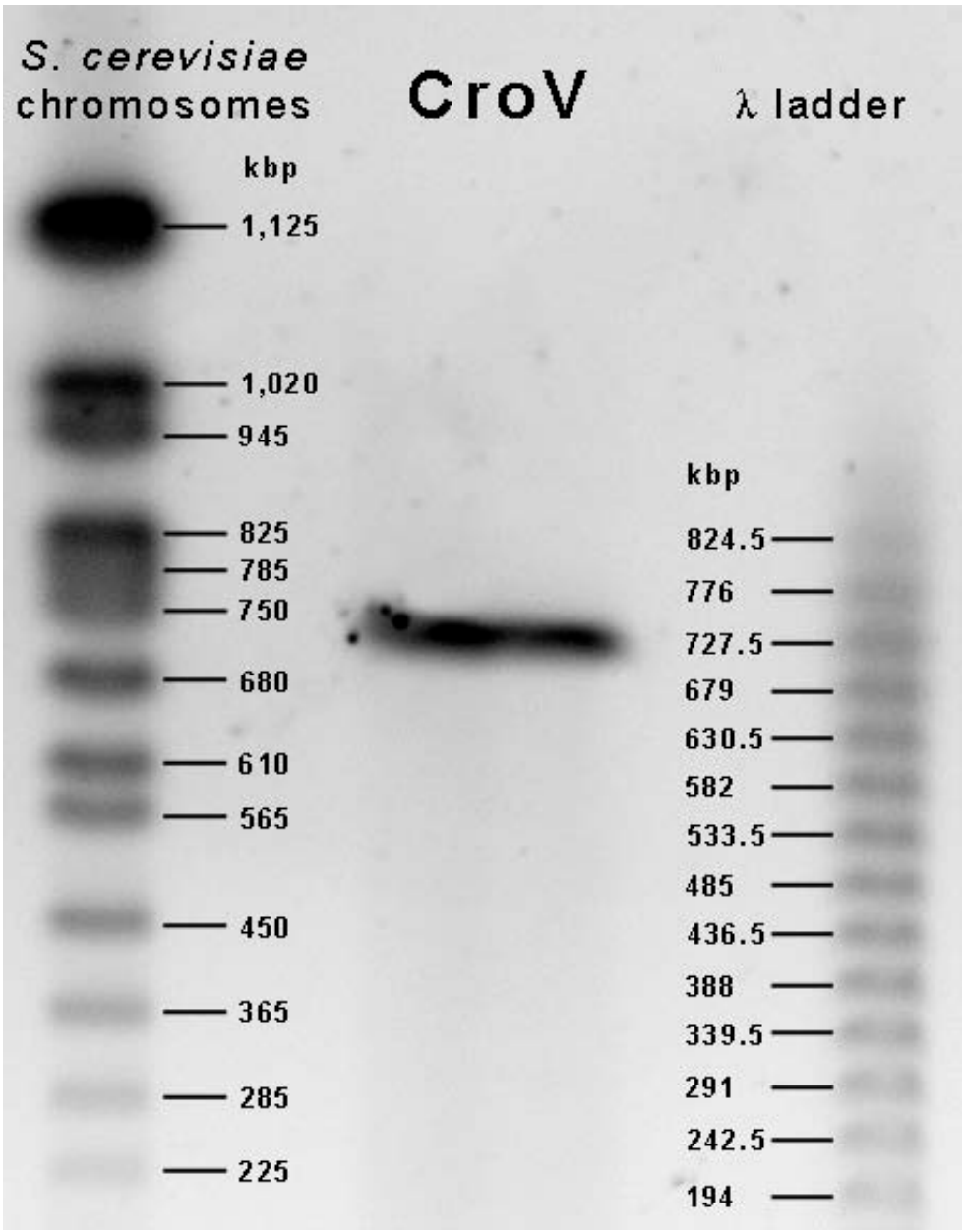


Figure 2.4 Pulsed-field gel electrophoresis of CroV restriction digests.

A, B, and C depict the same gel, which was stained and photographed after increasing electrophoresis times at 6 V/cm. DNA marker bands are labelled with their lengths in kpb. Individual DNA bands are labelled numerically. Identical bands have the same number. **(A)** Gel after 16 h with 0.1-8 sec ramped switch times. **(B)** Gel from (A) after additional 22 h with 8-25 sec ramped switch times. **(C)** Gel from (B) after additional 26 h with 25-150 sec ramped switch times.

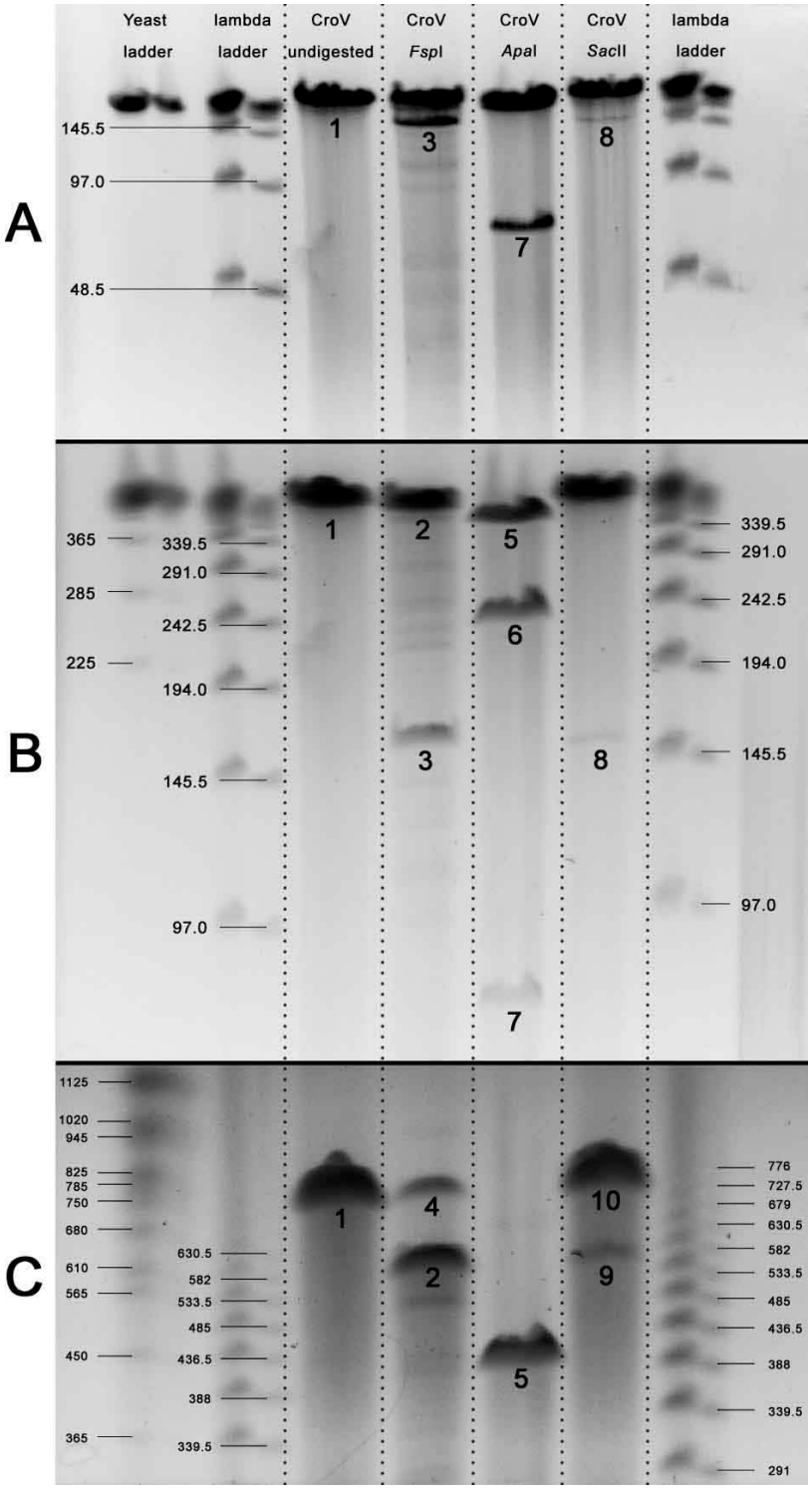


Table 2.1 Analysis of the PFGE restriction patterns from Fig. 2.4.

Restriction digest	Predicted cut sites	Observed number of restriction fragments	Bands as labelled in Fig. 2.4	Best size guess	Possible size range	Band intensity	Comments
Undigested	-	1	1	~750 kbp	730-800 kbp	very strong	too much DNA
***Fsp*I**	466,083	2	2	~600 kbp	575-620 kbp	strong	restriction fragment
			3	~168 kbp	155-175 kbp	strong	restriction fragment
			4	~750 kbp	730-770 kbp	weak	undigested
***Apa*I**	167,367 and 238,473	3	5	~425 kbp	400-440 kbp	strong	restriction fragment
			6	~245 kbp	240-250 kbp	strong	restriction fragment
			7	~71 kbp	65-75 kbp	strong	restriction fragment
***Sac*II**	468,677	2	8	~167 kbp	150-175 kbp	faint	restriction fragment
			9	~580 kbp	570-590 kbp	faint	restriction fragment
			10	~750 kbp	730-785 kbp	very strong	undigested

Figure 2.5 Genome diagram of CroV (next page).

Genome coordinates are given in kilobase pairs. Nested circles from outermost to innermost correspond to 1) predicted CDSs on forward strand and 2) reverse strand; 3) expression data for CDSs on forward strand and 4) reverse strand; 5) gene promoter type for CDSs on forward strand and 6) reverse strand; 7) location of repetitive DNA elements; 8) 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 kbp genomic segment of putative bacterial origin is shaded orange.

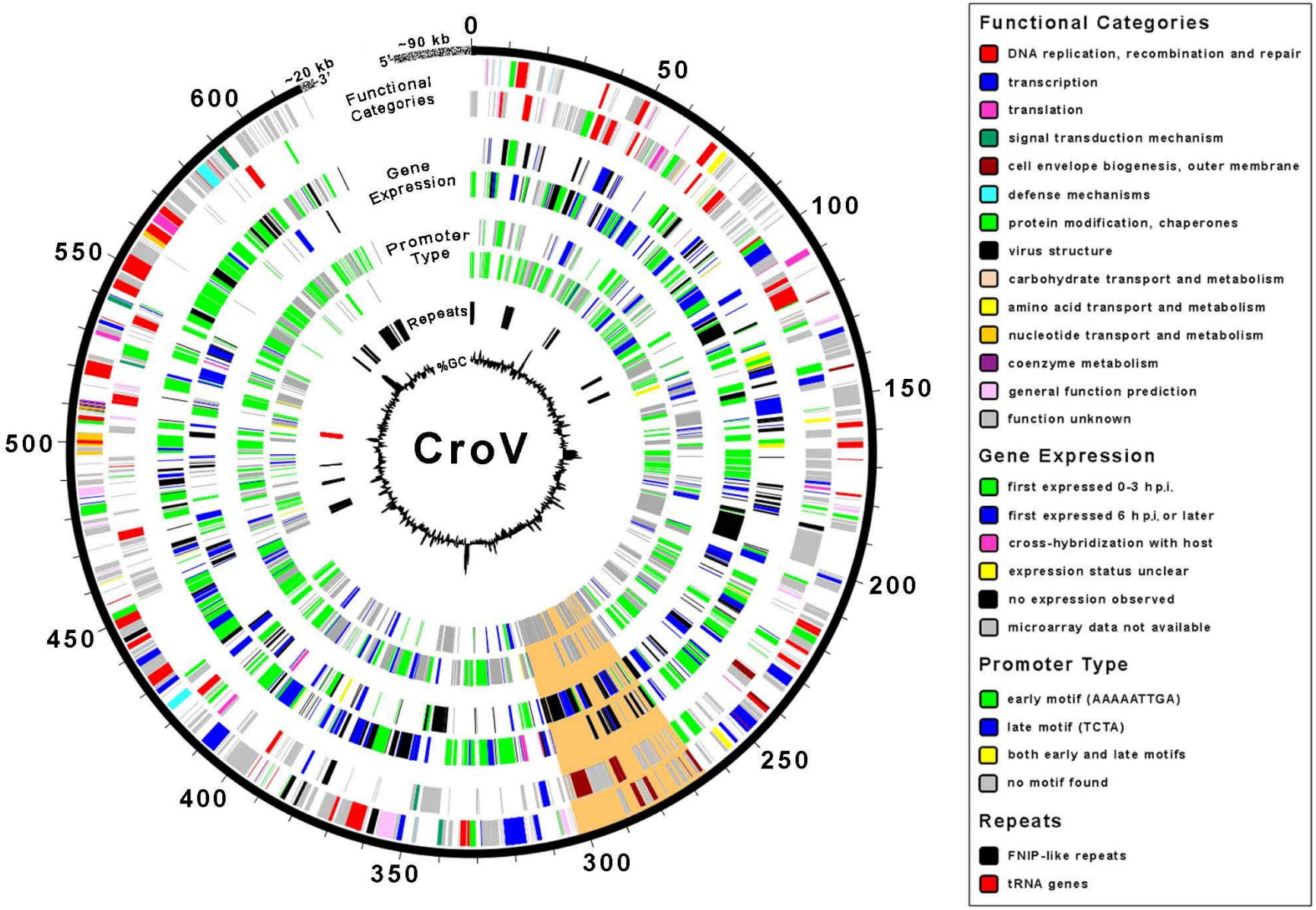
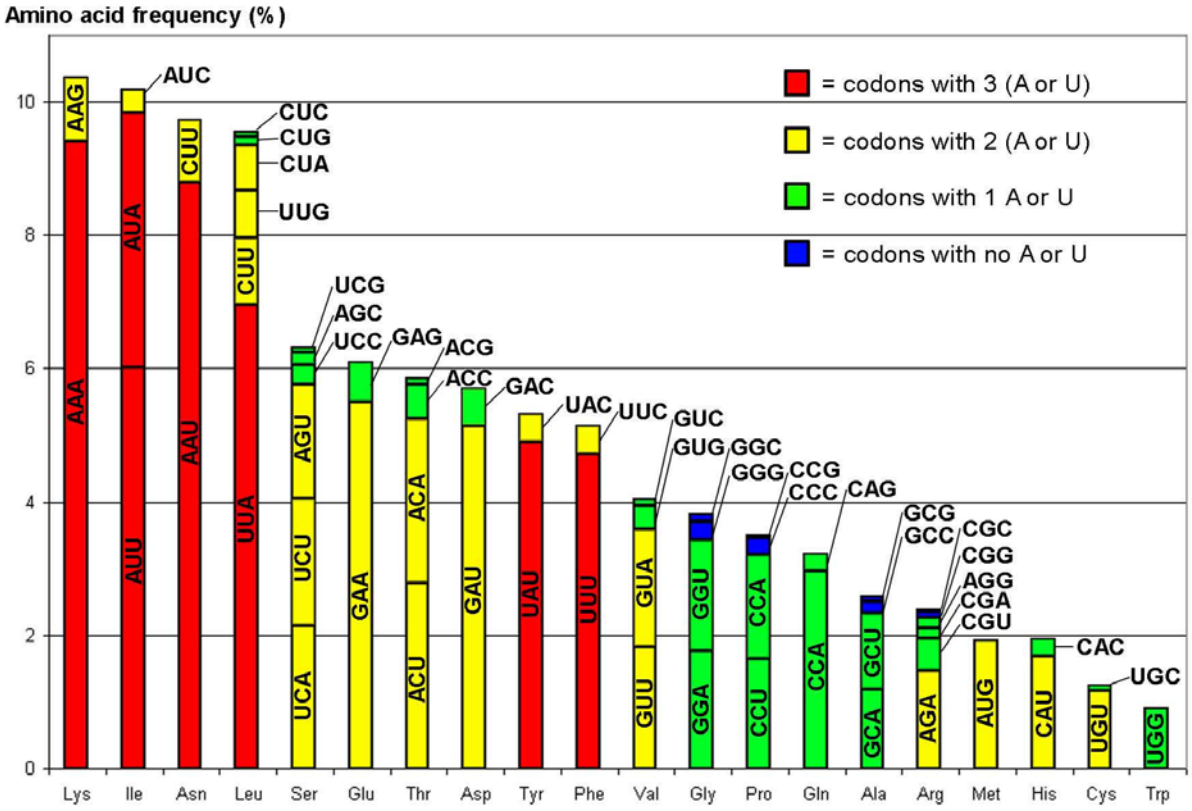

0
50
100
150
200
250
300
350
400
450
500
550
600
~90 kb
5'
~20 kb
3'
Functional Categories
Gene Expression
Promoter Type
Repeats
%GC
CroV
Functional Categories
DNA replication, recombination and repair
transcription
translation
signal transduction mechanism
cell envelope biogenesis, outer membrane
defense mechanisms
protein modification, chaperones
virus structure
carbohydrate transport and metabolism
amino acid transport and metabolism
nucleotide transport and metabolism
coenzyme metabolism
general function prediction
function unknown
Gene Expression
first expressed 0-3 h p.i.
first expressed 6 h p.i. or later
cross-hybridization with host
expression status unclear
no expression observed
microarray data not available
Promoter Type
early motif (AAAAATTGA)
late motif (TCTA)
both early and late motifs
no motif found
Repeats
FNIP-like repeats
tRNA genes

Figure 2.6 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 coloured 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.



Using conservative annotation criteria (see section 2.2.4), I identified 544 putative protein-coding sequences (CDSs) and 22 transfer RNA (tRNA) genes in the 618 kbp central region of the CroV genome, which has a coding density of 90.1%, or 0.88 CDSs per kb. The average CDS is 1025 nucleotides (nt) or 342 aa in length, and coding capacities range from 47 to 3,337 aa. The forward strand contains 290 CDSs and all 22 tRNA genes, the reverse strand encodes 254 CDSs. Applying a BLASTP E-value cutoff of 10^{-5}, 267 CDSs (49%) displayed similarity to sequences in GenBank and 134 CDSs (25%) were assigned to one or more Clusters of Orthologous Groups of proteins (COGs, E-value < 0.001) (Fig. 2.7). The most common COG categories are DNA replication & repair, protein modification & chaperones, and transcription. The least represented categories are energy production & conversion, cell division & chromosome partitioning, and signal transduction mechanisms. Given the parasitic lifestyle of viruses, this distribution is not surprising; however, it was unexpected to find 14 CDSs assigned to "translation, ribosomal structure and biogenesis". Appendix B contains a detailed list of all 544 CDSs, their physical properties and predicted functions.

Figure 2.7 Functional categories of Clusters of Orthologous Groups of proteins (COGs) identified in CroV.

Colour code: blue, information storage and processing; red, cellular processes; green, metabolism-related categories; black, poorly characterised categories.

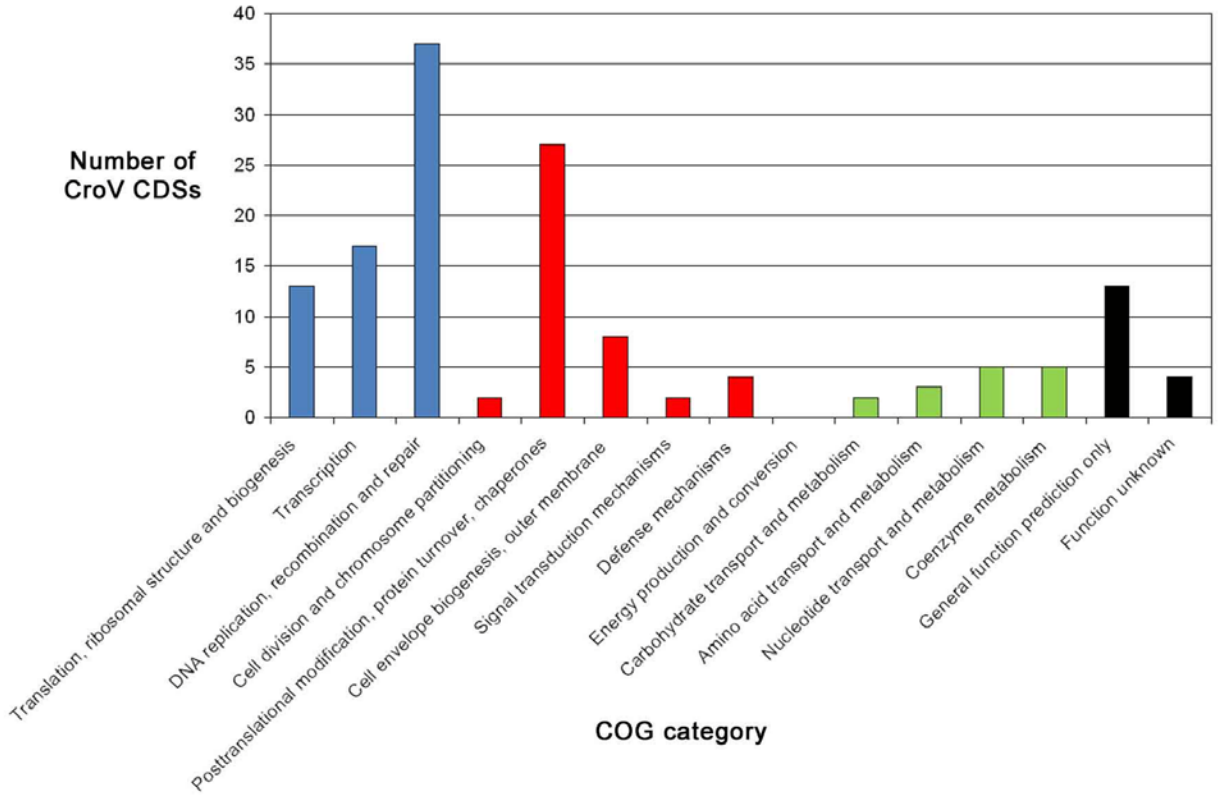
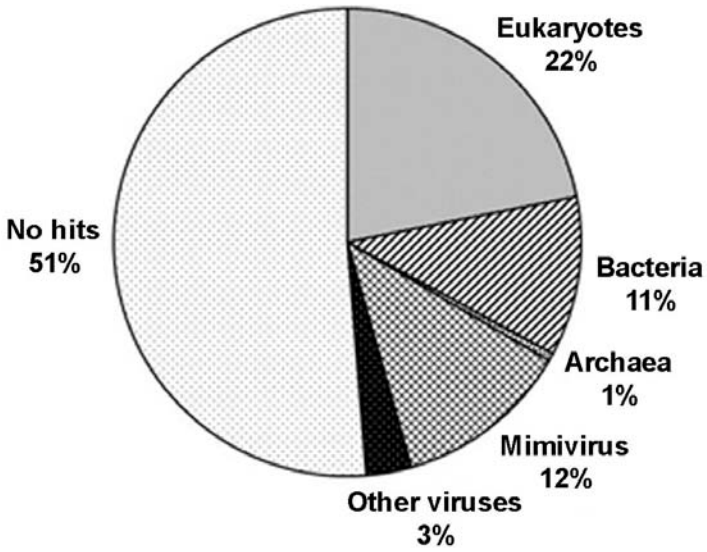


Figure 2.8 Distribution of top BLASTP hits for CroV CDSs.

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 10^{-5}.



Based on the distribution of top BLASTP hits, about half of the CroV genes displayed similarities to proteins found in eukaryotes, bacteria, archaea and other giant viruses (Fig. 2.8). 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 (Table 2.2). Owing to the large number of proteins found in CroV, only a fraction of them can be discussed within the scope of this thesis. The following sections provide an overview of the most important enzymatic functions encoded by this new giant virus.

Table 2.2 Novel viral features in the CroV genome.

Predicted Function	CroV CDS	Category
CPD class I photolyase	crov115	DNA replication and repair
Exodeoxyribonuclease VII large subunit	crov048	DNA replication and repair
Exonuclease III / AP endonuclease family 1	crov106	DNA replication and repair
DNA topoisomerase IB, human subfamily	crov152	DNA replication and repair / Transcription
Elp3-like histone acetyltransferase	crov391	Transcription
Eukaryotic translation initiation factor 2 alpha, SUI2 homologue	crov162	Translation
Eukaryotic translation initiation factor 2 gamma	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

2.3.2 Translation genes

Viruses rely primarily on the protein translation apparatus of their hosts; it is therefore unusual to find viral genes associated with protein synthesis. CroV encodes an isoleucyl-tRNA synthetase (IleRS) and putative homologues of eukaryotic translation initiation factors eIF-1, eIF-2α, eIF-2β/eIF-5, eIF-2γ, eIF-4AIII, eIF-4E, and eIF-5B. Initiation is the rate-limiting step of translation and is thus tightly regulated. During initiation, viral mRNAs compete with cellular transcripts for ribosome binding [162]. In fact, all of the CroV-encoded eIFs are involved in the formation of the 43S preinitiation complex, which consists of the 40S ribosomal subunit and eIFs 1, 1A, 2, 3, and 5 [163]. This multimeric complex binds to the 5'-cap structure of mRNAs via eIF-4E, eIF-4G, and eIF-4A and then scans the 5'-UTR for an AUG start codon, a process that is mediated by eIF-1. During the scanning phase, the initiator methionyl tRNA is bound to the GTP-bound form of the heterotrimer eIF-2 (composed of α, β and γ subunits). Experiments in yeast have shown that mutations in eIF-1 can lead to recognition of non-AUG start codons [164]. The CroV-encoded eIF-1 homologue may thus be capable of recognizing non-canonical start codons, which are postulated for 15 CroV CDSs (see section 2.2.4). The translation initiation factors found in CroV could thus potentially replace host factors in the 43S preinitiation complex and shift the specificity of mRNA recognition from host transcripts towards viral transcripts. A role of the CroV capping enzyme in the viral mRNA recognition process via CroV-specific 5' cap structures is also possible.

The evolutionary origin of the CroV translation-associated genes is not clear. Among these genes, the predicted IleRS encoded by crov505 showed the highest conservation to cellular homologues and was chosen for phylogenetic analysis. As shown in the phylogenetic tree in Fig. 2.9, the CroV IleRS homologue does not seem to be the result of horizontal gene transfer (HGT) from a eukaryotic host, because it branches basally to eukaryotes. Although viruses in general are known to have accelerated mutation rates, a recent study has shown that the evolutionary rates in three large dsDNA viruses are comparable to those in vertebrates [165]. Hence, it is unlikely that increased mutation rates account for the observed distance of CroV IleRS and eukaryotic homologues.

The CroV genome encodes 22 tRNA genes, as predicted by the transfer RNA gene prediction software tRNAScan-SE. The tRNA genes are clustered in a 2.8 kbp region around position 510,000 (Table 2.3, Figs. 2.5, 2.10) and have the following aa specificites: 9x Leu (TAA), 5x Ser (CGA), 3x Lys (TTT), 1x Tyr (GTA), 1x Asn (GTT), 1 ochre stop codon suppressor tRNA and 2 tRNAs of unknown specificity, one of which could be specific for isoleucine, containing a deletion in the anticodon loop (tRNA No. 9 in Table 2.3). The tRNA gene cluster seems to have

expanded by gene duplication, as some blocks of tRNA genes are present in multiple copies, such as the sequence Leu-Ser-Leu-Lys, which occurs three times in the tRNA gene cluster (Fig. 2.10, Table 2.3). The finding of an IleRS gene and 22 tRNA genes may reflect the importance of the respective aa in the viral infection cycle.

Figure 2.9 Phylogenetic analysis of isoleucyl-tRNA synthetases.

Bayesian Inference tree (unrooted) of a 431 aa alignment of isoleucyl-tRNA synthetases. Nodes with posterior probabilities of 0.90 or higher are marked with dots. Sequences are coloured orange for archaea, blue for bacteria, and green for eukaryotes. GenBank accession numbers are given for each sequence.*note: In ML analyses, CroV IleRS branched at the indicated position. The alignment was created by Mr. Shi Mang as part of an undergraduate project.

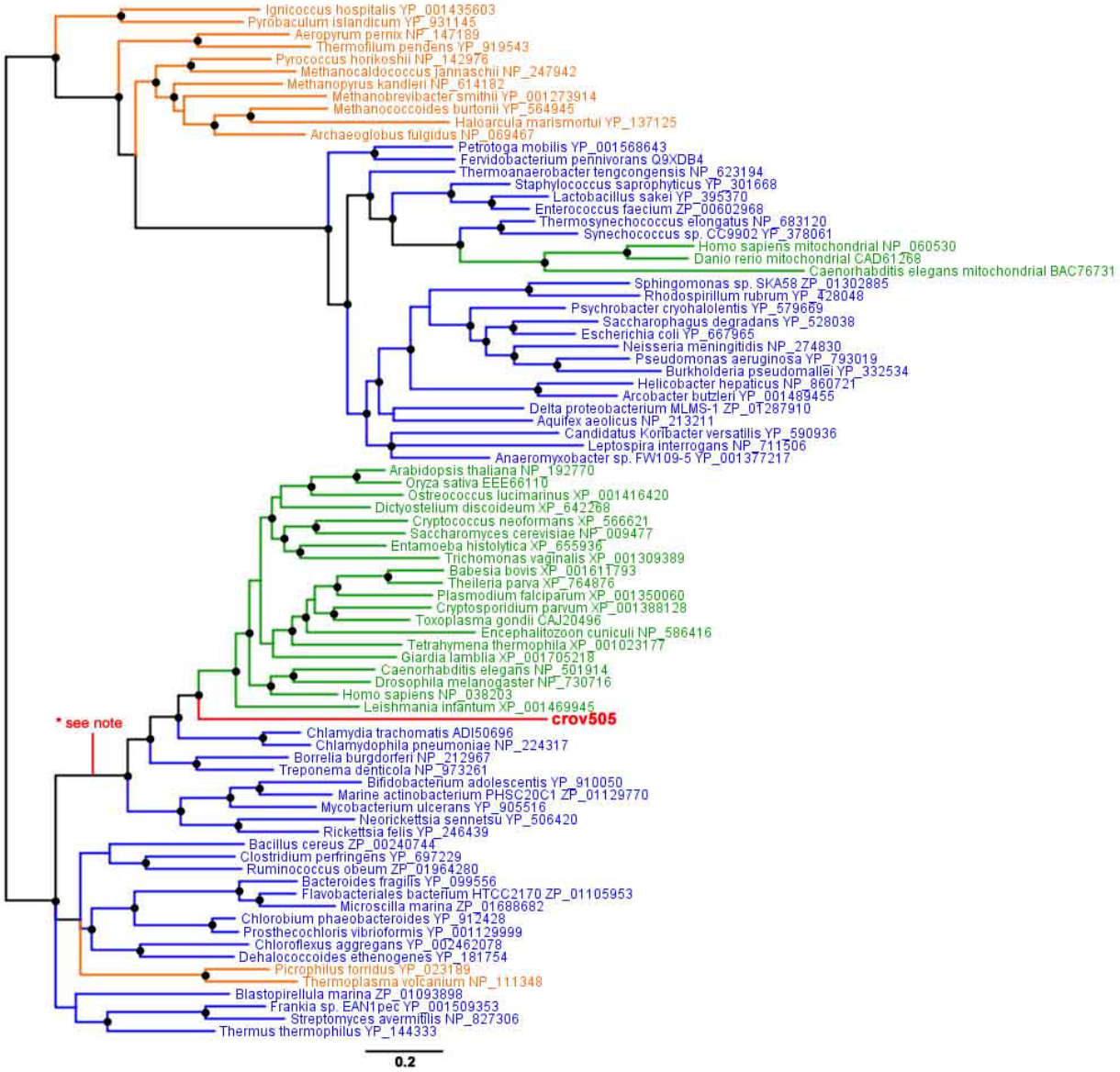


Figure 2.10 The CroV tRNA gene cluster.

This modified screenshot from the Artemis genome viewer [145] shows the genomic organization of the 22 tRNA genes. Individual tRNA genes are coloured according to their anticodon and match the colour code used in Table 2.3. The graph displays the GC content (window size: 100 bp) of this genomic region (6-51% G+C with a mean of 23% G+C). Predicted tRNA genes correspond to regions of high GC content.

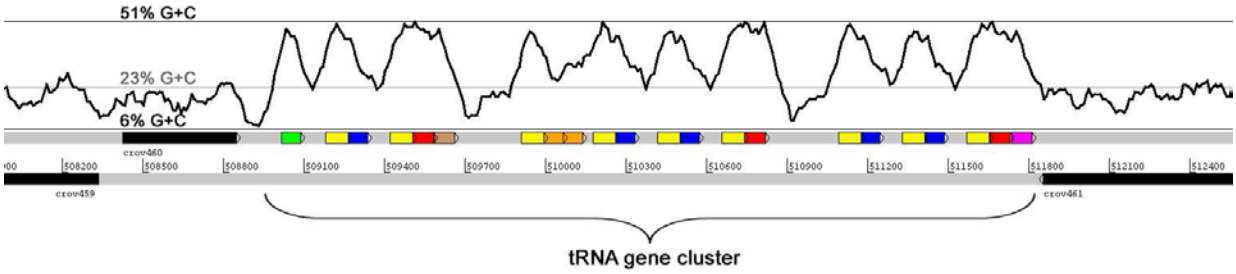


Table 2.3 Transfer RNA genes in the CroV genome.

tRNA genes sharing the same anticodon are shown in the same colour. The Cove Score reflects the statistical significance of the tRNA prediction and is dependent on the utilized covariance model.

tRNA No.	Begin	End	Type	Anticodon	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

Lysine, the most frequent aa in the CroV proteome (Fig. 2.6), is represented by three tRNAs; the second most frequent aa, isoleucine, has an associated tRNA synthetase and one possible tRNA gene; the third most frequent aa, asparagine, has one associated tRNA gene, and the fourth and fifth most frequent aa, leucine and serine, have nine and five associated tRNA genes, respectively. Although these results seem to imply a relationship between codon usage and translational genes in CroV, it is not known whether the tRNA genes are expressed or functional; thus, their anticodons could also be a result of mutational adaptation to the high AT content of the CroV genome. In Mimivirus, no correlation is observed between codon usage and virus-encoded aminoacyl-tRNA synthetases or tRNAs [59].

Further CroV CDSs with predicted functions in protein biosynthesis are two tRNA-modifying enzymes, tRNA pseudouridine 5S synthase and $tRNA^{Ile}$ lysidine synthetase. These genes add to a still exclusive, but rapidly growing number of virus-encoded protein translation components. Some tRNA genes are scattered among bacteriophages and eukaryotic viruses such as the phycodnaviruses [58, 166], and four tRNA synthetases along with several putative translation factors are found in Mimivirus [26]. 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 "life style" that is less dependent on host cell components than those of smaller viruses.

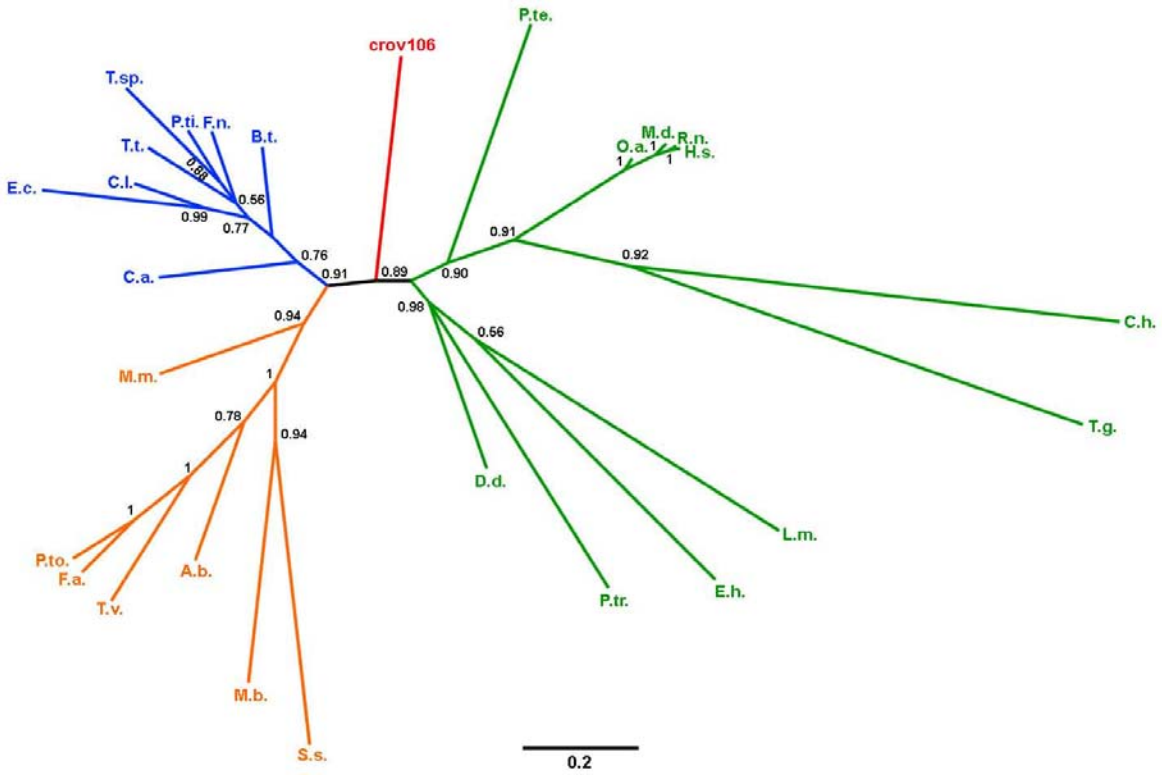
2.3.3 DNA repair genes

The ability to repair various kinds of DNA damage is well documented among large dsDNA viruses [167-169]. 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 several DNA repair proteins. Putative components of several DNA repair pathways were identified, including a presumably complete base excision repair (BER) pathway with formamidopyrimidine DNA glycosylase, a family 1 apurinic/apyrimidinic (AP) endonuclease, a family X DNA polymerase, and an NAD-dependent DNA ligase. Oxidative stress may lead to modification of DNA bases, which can result in mutations during DNA replication. The BER pathway is able to reverse these harmful modifications. The formamidopyrimidine DNA glycosylase could initiate BER by cleaving the N-glycosidic bond of oxidized purines, creating an AP site. AP endonuclease would then cleave the phosphodiester backbone at AP sites. A DNA polymerase such as the family B or family X enzymes encoded by CroV might fill in the resulting gap and the NAD-dependent ligase would seal the sugar-phosphate backbone. CroV may thus be able to catalyze the entire BER pathway by relying exclusively on enzymes encoded by its own genome.

The CroV AP endonuclease is the first viral homologue in this family and its phylogenetic position between eukaryotes and archaea/bacteria makes it unlikely that this gene has been recently acquired by HGT (Fig. 2.11).

Figure 2.11 Phylogenetic analysis of AP endonucleases.

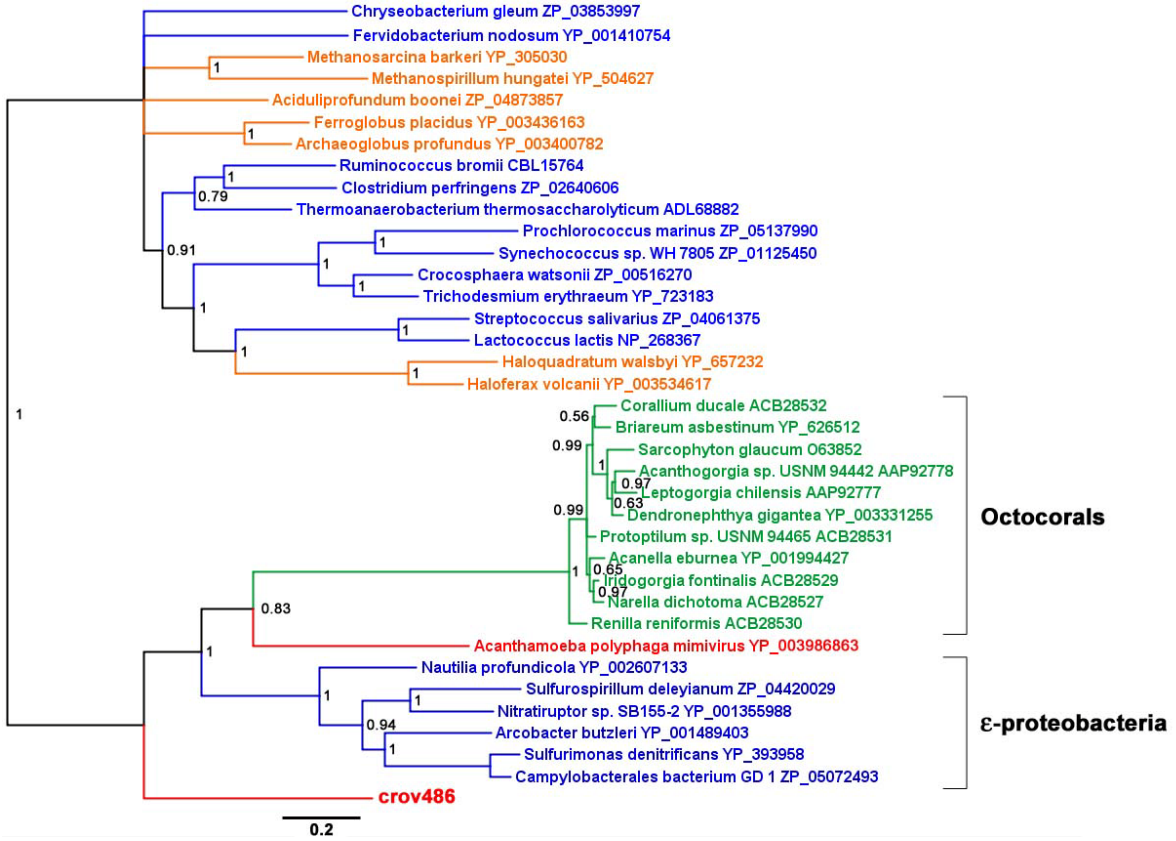
This Bayesian reconstruction of AP endonuclease phylogeny is based on a 168 aa alignment and is shown as a radial tree to emphasize the central branching point of the CroV protein. Branch nodes are labelled with posterior probabilities. Archaea are coloured orange, bacteria blue, and eukaryotes green. Species abbreviations and GenBank accession numbers are as follows: A.b., *Aciduliprofundum boonei*, ZP_04875824; B.t., *Bacillus thuringiensis*, ZP_04085756; C.a., *Coraliomargarita akajimensis*, YP_003548813; C.h., *Cryptosporidium hominis*, XP_666752; C.l., *Clostridium leptum*, ZP_02080964; crov106, ADO67139; D.d., *Dictyostelium discoideum*, XP_642518; E.c., *Eubacterium cylindroides*, CBK88844; E.h., *Entamoeba histolytica*, XP_650532; F.a., *Ferroplasma acidarmanus*, ZP_05570798; F.n., *Fusobacterium nucleatum*, ZP_00144662; H.s., *Homo sapiens*, NP_001632; L.m., *Leishmania major*, XP_001682147; M.b., *Methanosarcina barkeri*, YP_306703; M.d., *Monodelphis domestica*, XP_001379263; M.m., *Methanococcus maripaludis*, YP_001329475; O.a., *Ornithorhynchus anatinus*, XP_001516262; P.te., *Paramecium tetraurelia*, XP_001450332; P.ti., *Prevotella timonensis*, ZP_06288822; P.to., *Picrophilus torridus*, YP_023405; P.tr., *Pan troglodytes*, NP_001074954, *Phaeodactylum tricornutum*, XP_002181718; R.n., *Rattus norvegicus*, NP_077062; S.s., *Sulfolobus solfataricus*, NP_343662; T.g., *Toxoplasma gondii*, XP_002370968; T.sp., *Turicibacter sp.* PC909, ZP_06622751, T.t., *Thermoanaerobacter tengcongensis*, NP_623773; T.v., *Thermoplasma volcanium*, NP_110565.



Further DNA repair proteins in the CroV genome include DNA mismatch repair protein MutS, XPG endonuclease, a homologue of the alkylated DNA repair protein AlkB, and two DNA photolyases. The bacterial mismatch repair ATPase MutS (crov486) marks an interesting case, as it is only the second instance of a virally encoded MutS protein. The first viral MutS homologue was found in Mimivirus and turned out to be most closely related to a group of corals, the *Octocorallia*, who uniquely encode this gene in their mitochondrial genomes and probably acquired it by HGT from ε-proteobacteria [170, 171]. Although the CroV version is closely related to the Mimivirus MutS (BLASTP E-value 10^{-112}, 31% similarity over 1035 aa), phylogenetic reconstruction revealed that CroV MutS has a more basal branching position, as shown in the Bayesian Inference tree in Fig. 2.12. This suggests that a giant virus that was more closely related to Mimivirus than to CroV may have mediated the HGT from bacteria to corals [170]. However, Mimivirus is no longer the only virus encoding a MutS protein and giant viruses may represent a reservoir and, given the deep branching position of CroV MutS, perhaps even a source of cellular DNA repair genes.

Figure 2.12 Phylogenetic analysis of MutS ATPases.

Unrooted Bayesian Inference tree of a 374 aa alignment of MutS DNA mismatch repair proteins. Archaea are coloured orange, bacteria blue, and eukaryotes green. Nodes are labelled with posterior probabilities and GenBank accession numbers are given for each sequence.



The presence of two photolyase genes in the CroV genome is very unusual. Photolyases are classified into three major groups: cyclobutane pyrimidine dimer (CPD) photolyases, (6-4) photolyases, and single-stranded DNA photolyases ([172] 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 (Fig. 2.13). The gene product of crov115 is a predicted CPD class I photolyase and represents the first viral homologue in this class (Figs. 2.13, 2.14). 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 [172] (Figs. 2.13, 2.15). The only eukaryotic member in this group (*Paramecium tetraurelia*) is also the closest homologue to the CroV and Mimivirus sequences and may have acquired this gene by HGT from a giant virus (Fig. 2.15).

Figure 2.13 Phylogenetic analysis of the photolyase/chryptochrome family.
The unrooted Bayesian Inference tree of photolyases and chryptochromes is based on 73 conserved sites. Sequences are coloured orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labelled 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.

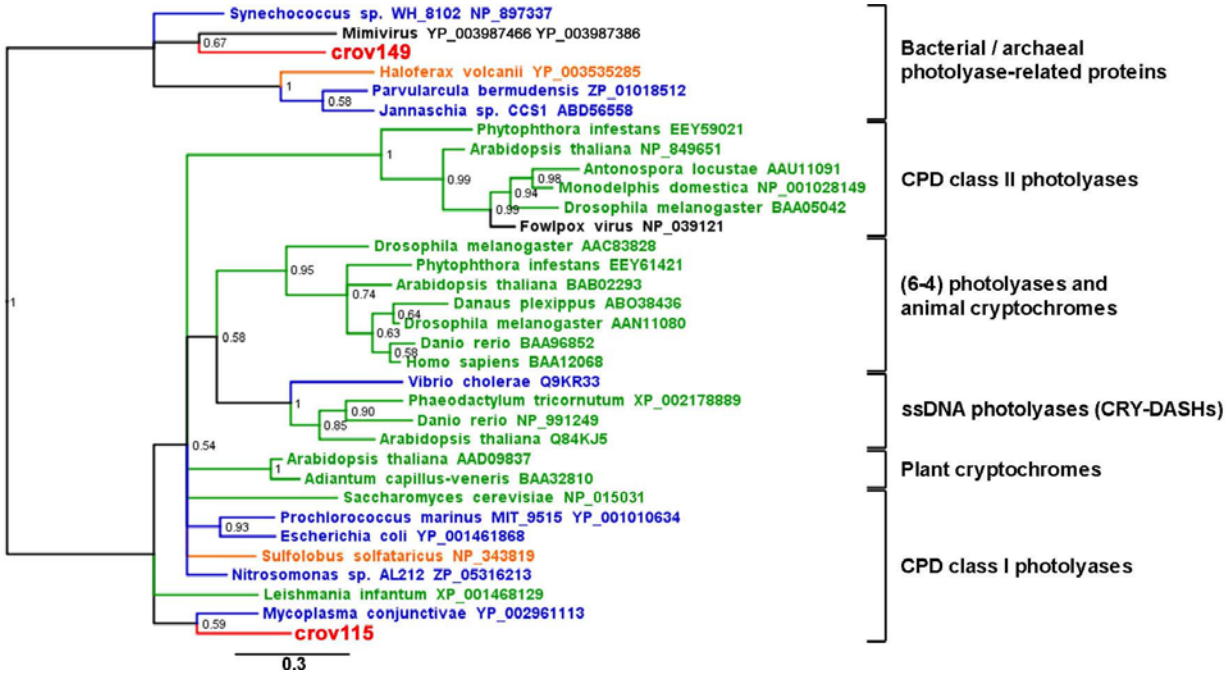


Figure 2.14 Phylogenetic analysis of CPD class I photolyases.

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

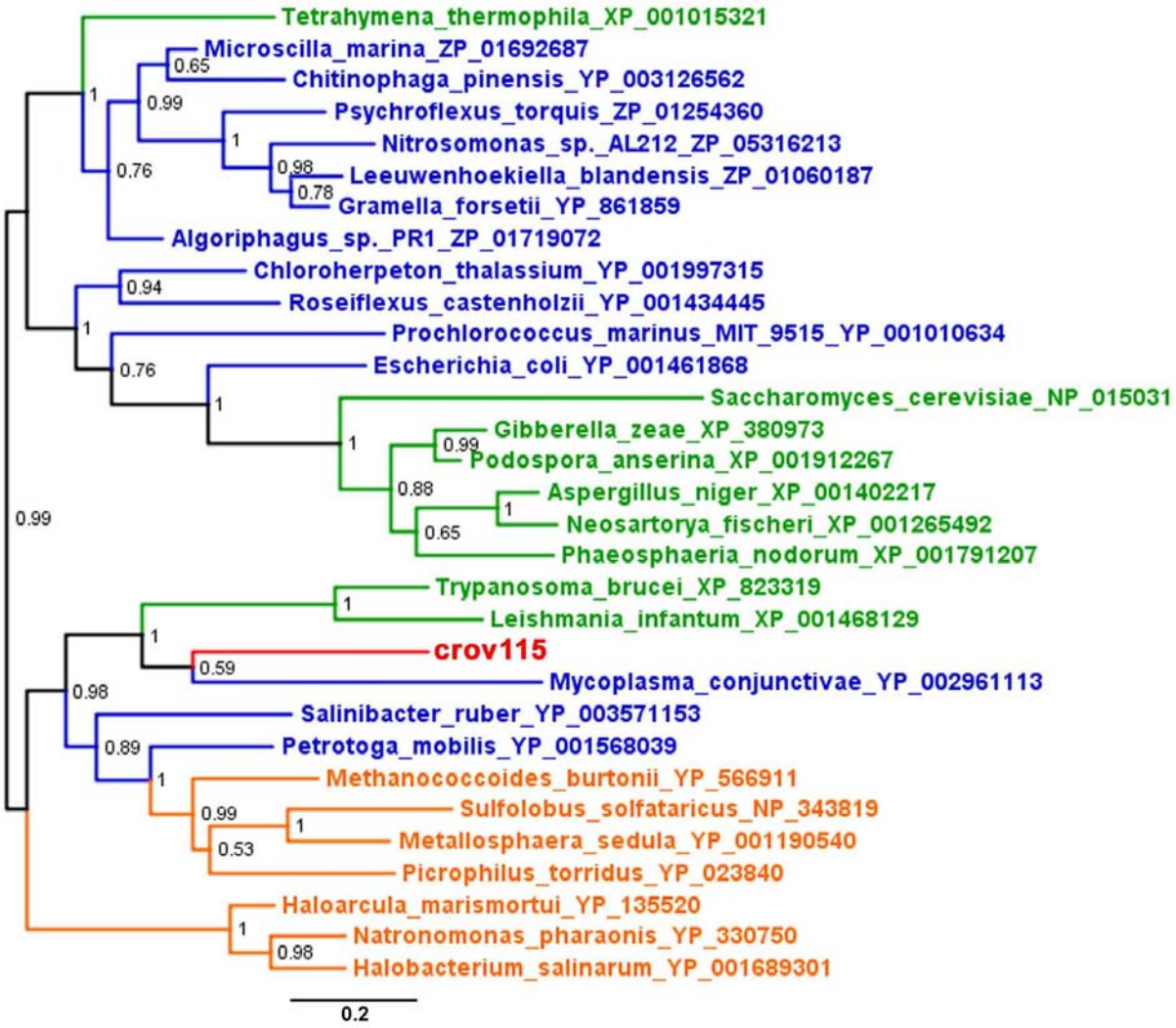
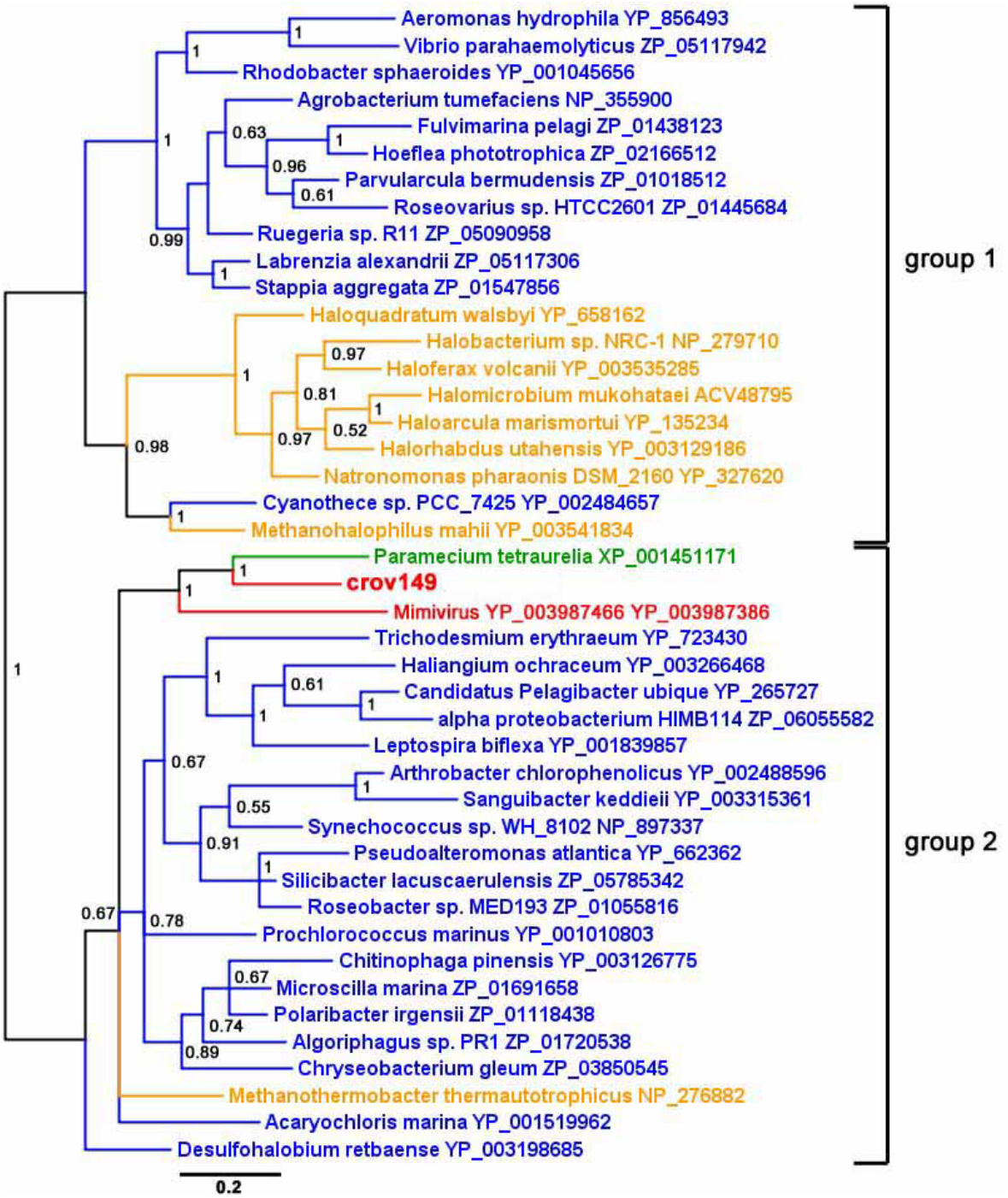


Figure 2.15 Phylogenetic analysis of the predicted DNA photolyase crov149.

The unrooted Bayesian Inference tree is based on a 157 aa alignment of DNA photolyases related to crov149. Sequences are coloured 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 two separate CDSs (R852/R853 and R855). Nodes are labelled with posterior probabilities and GenBank accession numbers are given for each sequence.



2.3.4 Transcription genes

Giant viruses typically encode 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 tri-functional 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 ELP3-like histone acetyltransferases (HATs, COG1243, E-value 10^{-46}), a gene previously not seen in viruses. Phylogenetic analysis indicates that eukaryotic HATs fall into two major groups, one being closer related to bacteria and containing mainly members of the chromalveolates and also some excavates, the other being closer to archaea and spanning various eukaryotic kingdoms (Fig. 2.16). The CroV homologue falls within the chromalveolate-excavate clade and may therefore be the result of a HGT event from a eukaryotic host. 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. Topoisomerases catalyze topological rearrangements of DNA molecules and are involved in DNA replication, recombination, and transcription. 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 [173]. CroV TopoIB is the first viral homologue of the eukaryotic subfamily, whereas the TopoIB homologue encoded by Mimivirus falls within the bacterial group (Fig. 2.17) and is functionally more similar to the poxvirus enzymes [80]. Despite apparently different evolutionary trajectories, the presence of three topoisomerase genes in CroV and Mimivirus suggests a crucial role for these enzymes in transcription, replication, or packaging of giant virus genomes.

Figure 2.16 Phylogenetic analysis of ELP3-like histone acetyltransferases.

Shown is an unrooted Bayesian Inference tree based on a 301 aa alignment of ELP3-like HAT homologues. Sequences are coloured orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labelled with posterior probabilities and GenBank accession numbers are given for each sequence.

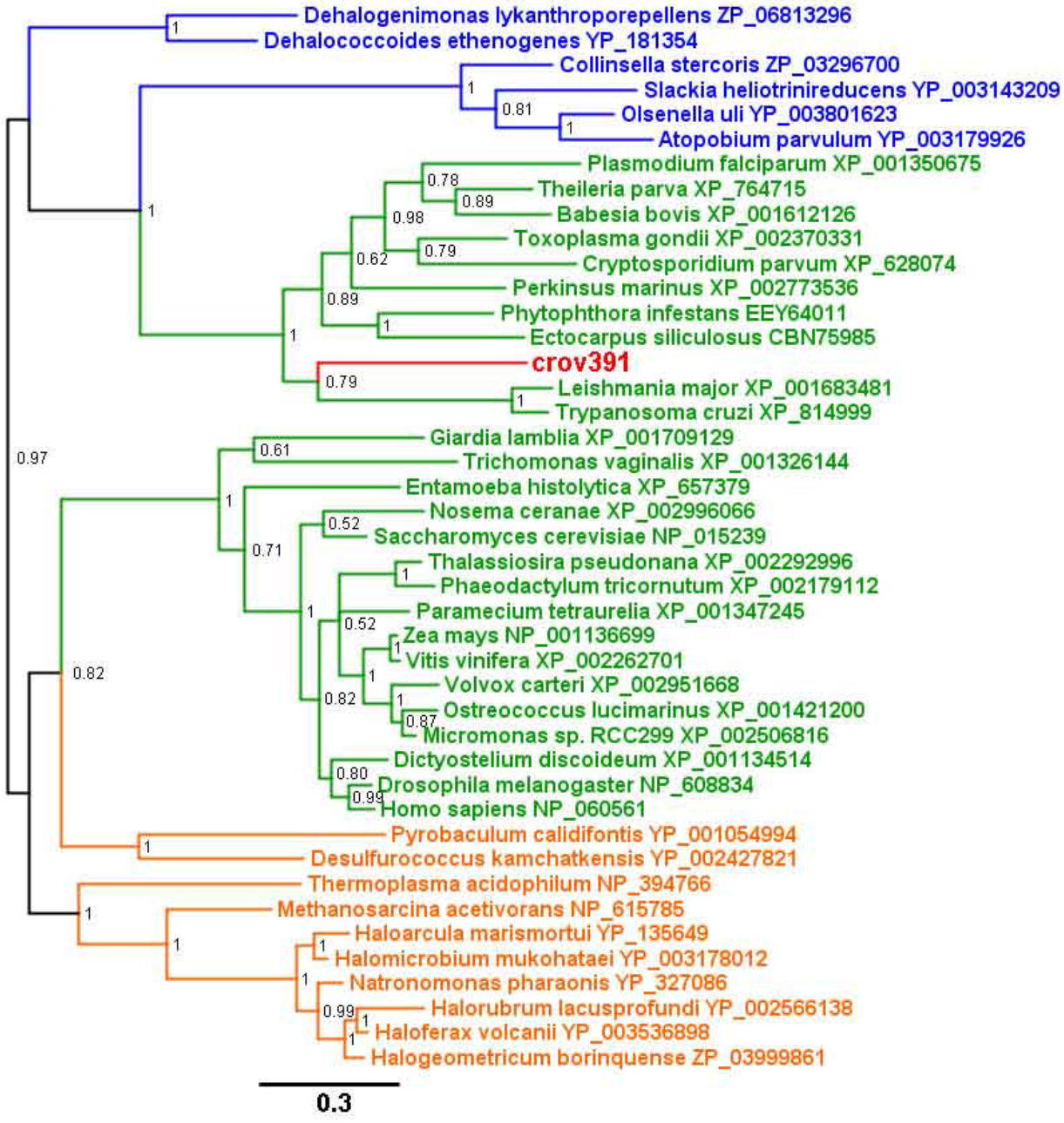
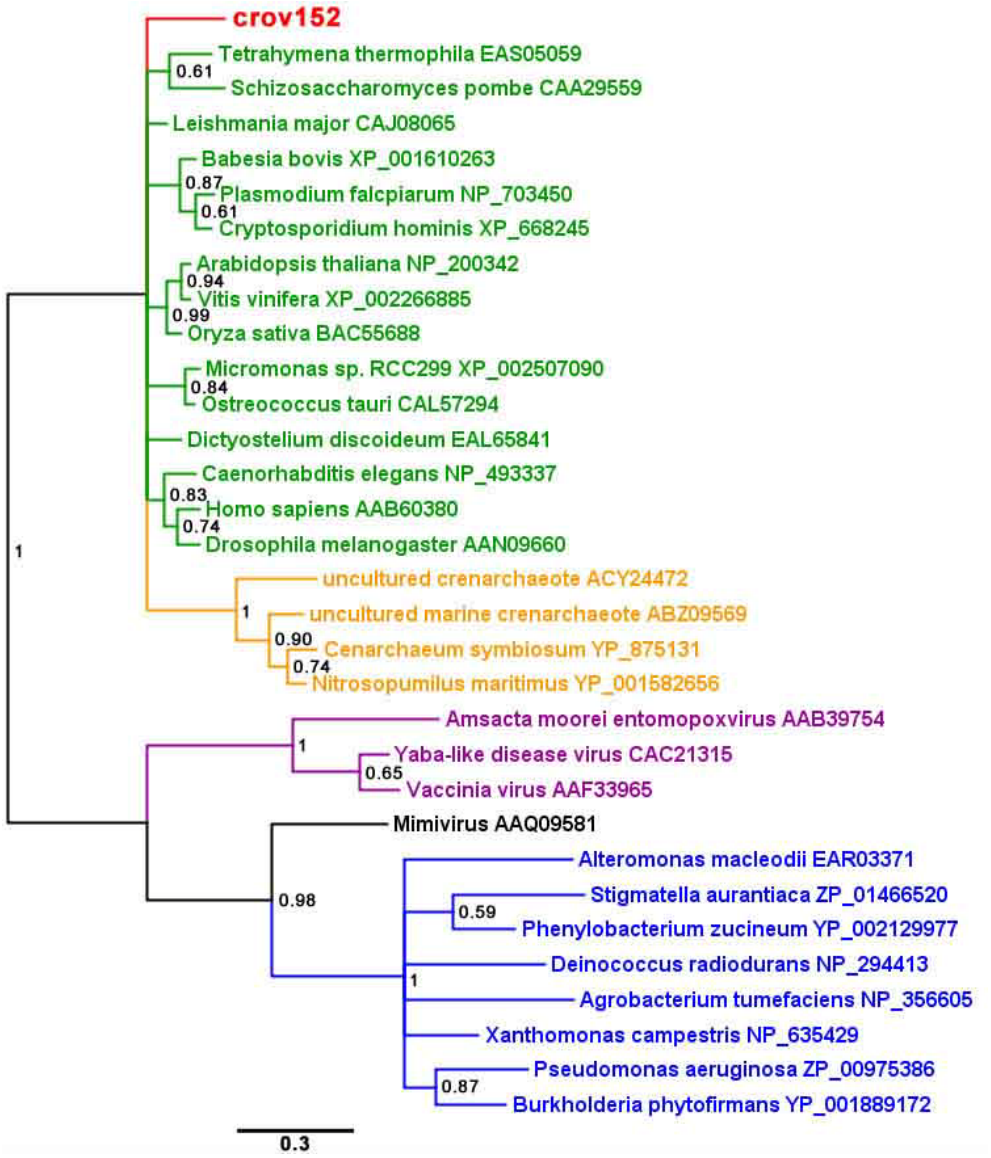


Figure 2.17 Phylogenetic analysis of type IB DNA topoisomerases.

The unrooted Bayesian Inference tree is based on a 66 aa alignment of DNA toposiomerases of type IB. Sequences are coloured orange for archaea, blue for bacteria, green for eukaryotes and purple for poxviruses. Nodes are labelled with posterior probabilities and GenBank accession numbers are given for each sequence.



2.3.5 DNA replication enzymes

Similarly to other members of the NCLDV clade that replicate in the host cytoplasm, CroV encodes several DNA replication enzymes. The B-family DNA-dependent DNA polymerase (PolB) is the main processing polymerase and is essential for NCLDV genome replication. These enzymes are highly conserved in all lineages and display a low propensity for HGT, which makes them an excellent target for phylogenetic reconstruction analyses. PolB is

therefore the most frequently used marker gene to infer evolutionary relationships among dsDNA viruses [99, 174]. The phylogenetic position of CroV PolB is discussed in section 2.3.13. In CroV, PolB is accompanied by the polymerase processivity factor PCNA (proliferating cell nuclear antigen) and other predicted replication proteins including a VV D5-type primase-helicase, four replication factor C subunits, a lambda-type exonuclease, at least eight DNA helicases, a DNA polymerase III subunit, and a DNA ligase. Two different types of DNA ligases are found in large DNA viruses, ATP-dependent and NAD-dependent DNA ligases. These two versions are mutually exclusive, i.e. the two types have never been found together within the same genome. The distribution pattern of DNA ligases in NCLDV is somewhat patchy, with ATP-dependent DNA ligases occurring in phycodnaviruses, ASFV, and chordopoxviruses, whereas NAD-dependent DNA ligases are found in mimiviruses, iridoviruses, and entomopoxviruses [50]. Yutin *et al*. [175] suggested that the common NCLDV ancestor encoded an NAD-dependent ligase from a bacteriophage source that was replaced by an ATP-dependent DNA ligase in some viral lineages and other lineages lost their DNA ligase entirely. The DNA ligase in CroV belongs to the NAD-dependent type, which may suggest that the CroV version is directly derived from the NCLDV ancestor.

Several phycodnaviruses have methylated bases in their genomes [176]. The extent to which bases in the CroV genome may be modified is unknown, but the presence of six predicted DNA methyltransferases (crov015, crov074, crov347, crov396, crov447, crov472) suggests that DNA methylation also occurs in CroV.

Replication of a large DNA molecule and packaging of the genome into a relatively small capsid causes topological problems that the virus has to resolve. In addition to the previously mentioned Topo genes, CroV encodes a predicted Holliday junction resolvase and a transposase-resolvase gene pair (crov356, crov357) similar to bacterial insertion sequences (IS elements), which might be used to overcome topological constraints.

2.3.6 Ubiquitin pathway components

Ubiquitin (Ub) is a 76-aa polypeptide that is extremely conserved among eukaryotes. Posttranslational modification by addition of Ub moieties can alter the function of a protein or target it for degradation [177]. This process involves three types of enzymes. First, a Ub-activating enzyme E1 transfers a Ub molecule to a Ub-conjugating enzyme E2. The E2 enzyme then either ubiquitinates the target protein directly, or passes the Ub further on to a Ub ligase E3, which in turn ubiquitinates the final substrate. Ub signalling appears to be a general strategy employed by NCLDVs to counter host defences, as multiple Ub-conjugating and Ub-hydrolysing

enzymes have been found in these viruses [50]. Furthermore, it has been shown that orthopoxvirus replication requires a functional Ub-proteasome system [178]. CroV contains a small arsenal of genes encoding proteins predicted to function in the Ub pathway, including the first viral homologue of an E1 Ub-activating enzyme, as well as six E2 Ub-conjugating enzymes, two deubiquinating enzymes, and one Ub gene. Interestingly, the CroV-encoded Ub is only 75 aa long and is missing the C-terminal glycine residue, which is the residue that is usually conjugated to substrate lysine residues. The specific means of how CroV and other giant viruses use Ub signalling to interact with their hosts remain to be determined.

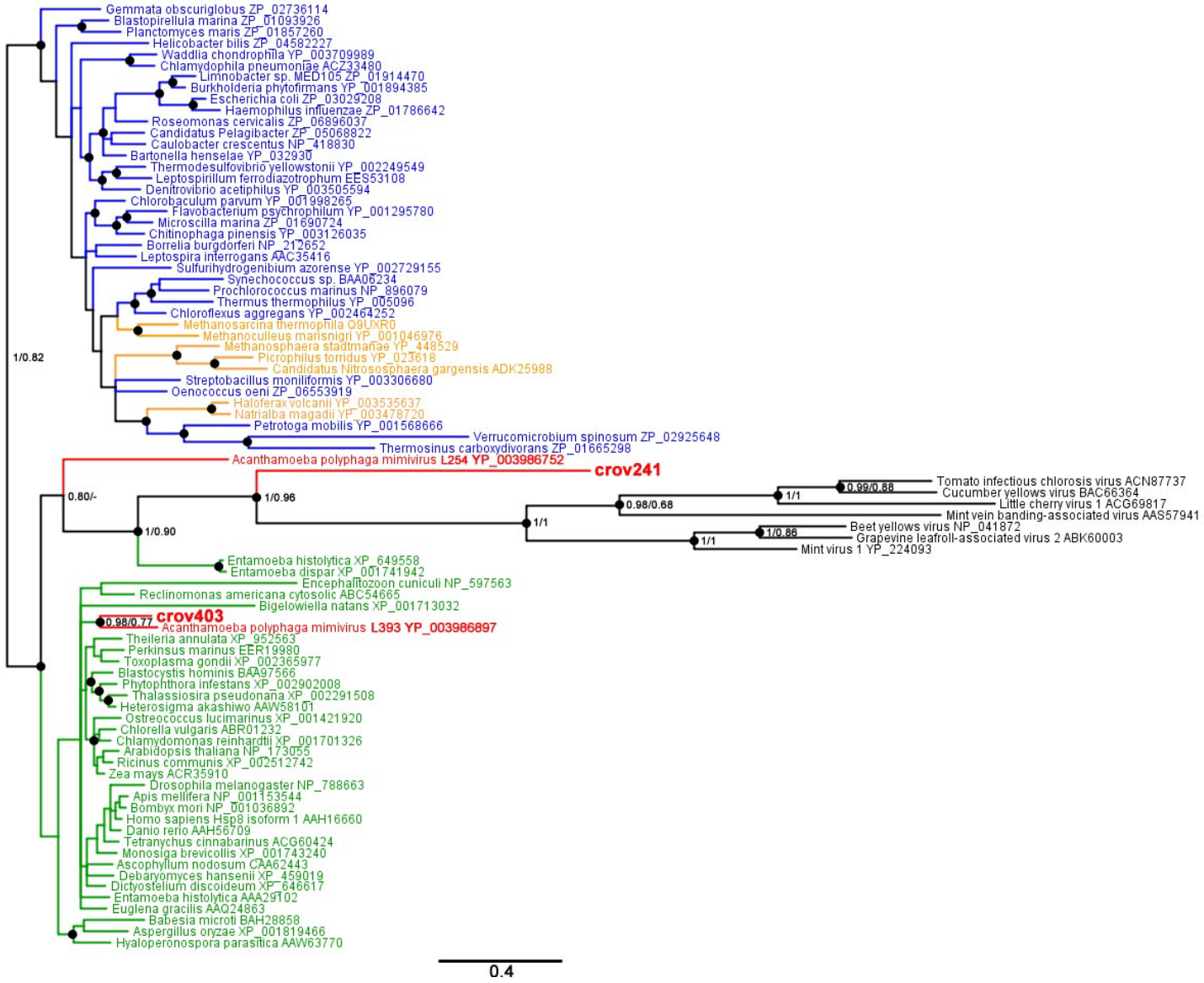
2.3.7 Molecular chaperones and redox proteins

Protein function is dependent on a correct three-dimensional structure. *In vivo* folding of nascent polypeptides is therefore often assisted by molecular chaperones such as heat shock proteins (HSPs). CroV was found to encode several proteins that were similar to cellular HSPs and other chaperones. Four CDSs contained domains similar to the HSP40/DnaJ family (crov127, crov170, crov210, crov375), two CDSs encoded HSP70/DnaK homologues (crov241, crov403), and three CDSs encoded predicted Lon proteases (crov036, crov237, crov478). Proteins of the HSP70 family have a ubiquitous distribution and have been used to reconstruct large-scale phylogenies [179]. Viral HSP70 homologues are rare and occur only in Mimivirus and in the *Closteroviridae*, a family of large positive-sense single-stranded RNA plant viruses [26, 180, 181]. Besides its traditional role in protein folding, HSP70 serves various other functions; e.g. in closteroviruses, it is involved in cell-to-cell virion movement [180]. HSP70 has also been shown to suppress apoptosis by inhibiting the recruitment of caspases to the apoptosome [182], a pathway that might be useful for viruses to exploit in order to prevent premature apoptotic cell death. The function of HSPs encoded by giant viruses is unknown; they could be involved in the folding of other viral proteins or in virion assembly, the latter condition has been experimentally proven for the closterovirus *Citrus tristeza* virus [183].

Phylogenetic analysis of HSP70/DnaK proteins from all three domains of life showed that CroV and Mimivirus encode two distinct versions of HSP70s (Fig. 2.18). The HSP70 homologue encoded by crov403 was most closely related to the L393 protein of Mimivirus and these two viral sequences grouped unambiguously with the eukaryotic HSP70 proteins, suggesting that the common ancestor of CroV and Mimivirus acquired this gene from a eukaryotic host by HGT. The crov241-encoded HSP70 homologue, on the other hand, was only distantly related to cellular HSP70s and was more similar to the closteroviral proteins. The second Mimivirus HSP70 homologue, L254, also fell in the closterovirus cluster, although it branched even deeper than crov241. Interestingly, some *Entamoeba* HSP70s were located on this branch, too.

Figure 2.18 Phylogenetic analysis of HSP70 proteins.

The Bayesian Inference tree is based on a 250 aa alignment of conserved regions of HSP70/DnaK proteins. As eukaryotes encode multiple HSP70 versions that are located in specific cell compartments, organellar homologues were not included in order to simplify the tree. Sequences are coloured orange for archaea, blue for bacteria, green for eukaryotes, black for closteroviruses, and red for giant viruses. Selected nodes are labelled with posterior probabilities and ML bootstrap values (500 replicates), all nodes with posterior probabilities of 0.90 of higher are marked with dots. GenBank accession numbers are given.



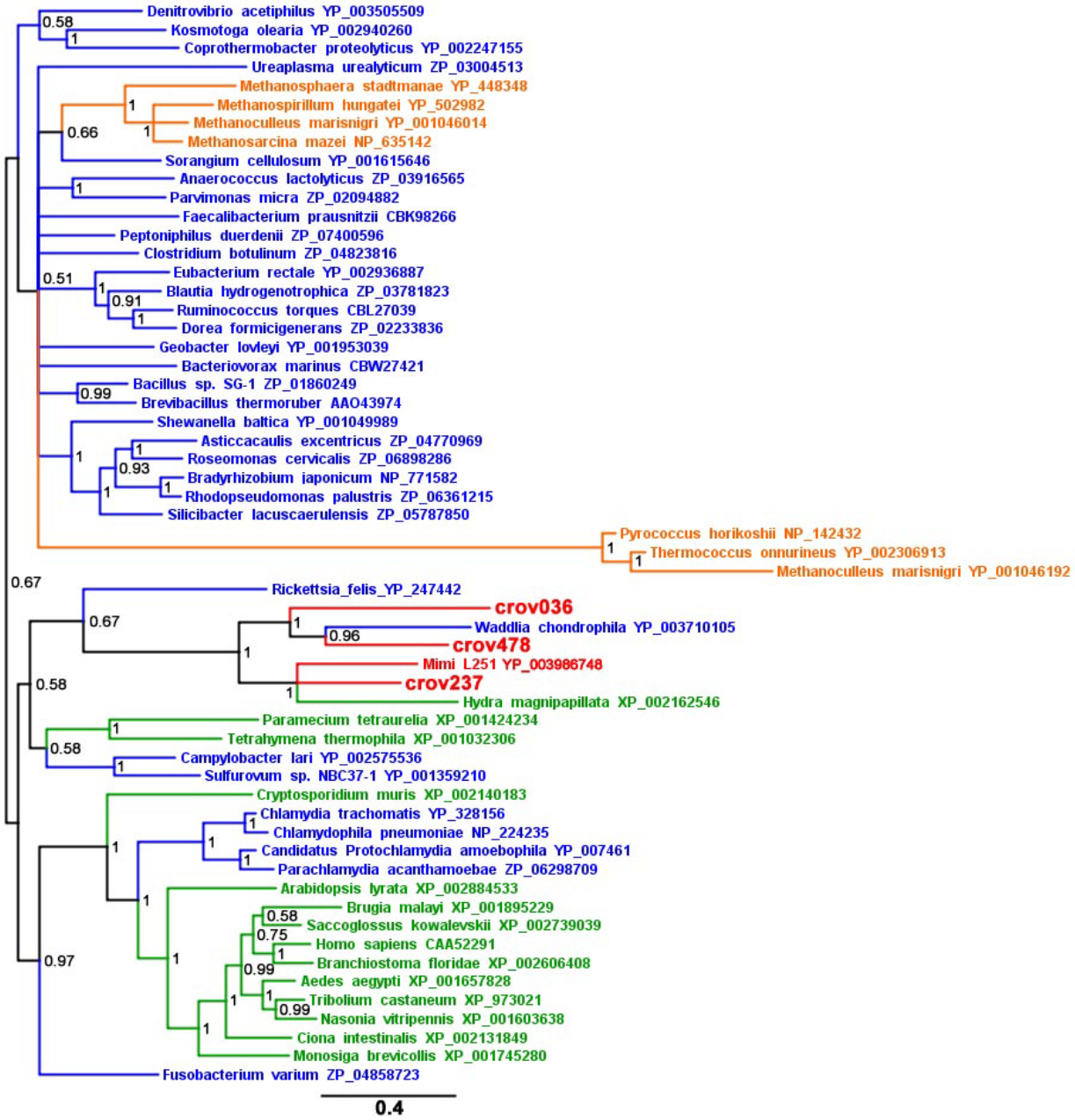
The closteroviral HSP70 genes may have found their way into the viral ssRNA genome by recombination with host plant mRNA [184], but their relationship to protist-infecting viruses with large dsDNA genomes remains enigmatic. It cannot be excluded that the observed clustering of obviously divergent HSP70 homologues may be the result of long-branch attraction, an artefact of phylogenetic analysis [185]. In order to allow molecular chaperones such as HSP70 to identifiy their substrates, co-chaperones are often required to mediate specific interactions with substrate proteins. The tetratrico peptide repeat (TPR) domain has been shown to act as a protein-protein interaction module, e.g. as a co-chaperone of HSP70 and HSP90 [186]. TPR domains are found in a variety of proteins that are usually part of multi-protein complexes and are involved in diverse cellular processes [187]. CroV encodes at least six TPR-containing proteins, which may indicate the importance of multi-protein complexes for CroV infection.

The ATP-dependent Lon proteases are part of the heat shock response in bacteria and are involved in a variety of cellular processes, but their main function is assumed to be the degradation of misfolded or oxidized proteins by recognizing hydrophobic regions of polypeptides, as they frequently occur in abnormal proteins [188]. The Lon proteases consist of an N-terminal domain of variable length responsible for substrate specificity, a central ATPase domain, and a C-terminal proteolytic domain. In eukaryotes, Lon is localized in the mitochondrial matrix where it degrades oxidatively modified proteins [189]. The first viral homologue of a Lon protease was found in Mimivirus [26] and CroV encodes three of these enzymes. Phylogenetic analysis showed that CroV and Mimivirus Lon proteases were closely related to each other and formed a distinct clade which also included Lon homologues from the mitochondrion of *Hydra magnipapillata* and the chlamydia bacterium *Waddlia chondrophila* (Fig. 2.19).

Some proteins require intra- or intermolecular disulfide bonds for structural integrity or enzymatic activity. The correct formation of these S-S bonds by oxidative protein folding is catalyzed by a family of molecular chaperones called protein disulfide isomerases (PDIs). PDIs can form, dissolve, and isomerize disulfide bonds, and are also able to bind to partially unfolded polypeptides [190, 191]. The two predicted PDIs and other redox proteins encoded by CroV are discussed in more detail in section 3.3.5.

Figure 2.19 Phylogenetic analysis of ATP-dependent Lon proteases.

The Bayesian Inference tree is based on a 354 aa alignment of conserved regions of Lon proteases. Lon homologues from *Hydra magnipapillata*, *Bruglia malayi*, and *Saccoglossus kowalevskii* are encoded by mitochondrial DNA. Sequences are coloured orange for archaea, blue for bacteria, green for eukaryotes, and red for viruses. Selected nodes are labelled with posterior probabilities and GenBank accession numbers are given for each sequence.



2.3.8 Inteins

No introns were detected in the genome, but four CroV CDSs contain an intein element. Inteins are self-splicing protein sequences inserted in the coding region of another protein, the host protein or "extein" [192]. Inteins have been termed parasitic genetic elements, because they are of no apparent benefit to their host. The intein element is transcribed and translated together with the host protein. It then removes itself from the host protein through autocatalytic splicing, covalently joining the N-terminal extein part with the C-terminal extein part to yield a functional full-length host protein. As of November 4, 2010, the New England Biolabs Intein Registry[10] listed 581 inteins from at least 250 different species [193]. Curiously, inteins are preferentially found in proteins associated with DNA replication and repair [194]. The insertion site in the host protein is highly conserved, such as the YGD/TDS catalytic motif in DNA polymerases. The assignment of inteins into allele groups is based on the observation that inteins sharing the same insertion site in orthologous host proteins are more similar to each other than to inteins found at different insertion sites in the same host protein; the former ones are thus termed intein alleles [195]. Inteins come in two flavours, large and mini [194]. Large inteins are equipped with a homing endonuclease (HE) domain, which enables the intein to invade empty target sites in other genes, species, or populations via horizontal transmission. Mini inteins, on the other hand, have lost their endonuclease domain and are therefore trapped in their current location.

All four CroV inteins are located in NCLDV core genes that are thought to play a key role in DNA replication and transcription (Fig. 2.20): PolB, TopoIIA, DNA-dependent RNA polymerase II subunit 2 (RPB2), and the large subunit of ribonucleotide reductase (RNR). With the exception of the RPB2 mini intein, all CroV inteins contain an HE domain and are thus considered large inteins. However, the HE motifs in PolB and RNR are hardly recognizable and may indicate that these HE domains are degenerating. The specific properties of the four inteins are listed in Table 2.4. Ten other inteins have been found in viruses infecting eukaryotes [193], including RNR inteins in four iridoviruses and *Paramecium bursaria* chlorella virus (PBCV) NY-2A [38]. PolB inteins inserted in the YGD/TDS catalytic site are present in several giant viruses, such as Mimivirus [196], *Heterosigma akashiwo* virus [197], and *Chrysochromulina ericina* virus [198]. A PCR-based study of phycodnaviral PolB sequences in Hawaiian coastal waters suggests that inteins may be very common in large marine DNA viruses [199]. With the exception of a gene fragment from *Emiliania huxleyi* virus 163 [200], the CroV RPB2 intein constitutes the only viral report of an intein in RPB2 and represents a new insertion site in this gene. Finally, the CroV TopoIIA intein is the first case of an intein in a DNA topoisomerase II

[10] http://www.neb.com/neb/inteins.html

gene, thus extending the known range of intein-containing genes. As indicated in Fig. 2.20, 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) [192]. It can therefore be hypothesized that the CroV inteins are capable of autocatalytic excision (for further discussion of the Rpb2 intein see section 3.3.3). Whereas the protein-splicing mechanism of inteins is well characterized and is broadly applied in biotechnology [201], the biological role and evolutionary origin of these genetic elements are not well understood. Although large inteins carry an HE domain and are potentially mobile, horizontal transmission of inteins is considered a rare event, as closely related species of an intein-containing species are often intein-free [202]. Inteins occur in all three domains of life and are extremely divergent; they are thus thought to be very ancient, probably dating back to the last universal common ancestor (LUCA) of eukaryotes, bacteria, and archaea [203]. Nagasaki *et al*. suggested that large dsDNA viruses might be a reservoir for inteins and could facilitate their horizontal transmission between different hosts [197]. The discovery of four inteins in CroV is in concordance with this idea.

Figure 2.20 Location of CroV inteins in their host proteins.

The linear peptide chains of the four intein-containing genes are drawn to scale. Regions shaded blue correspond to exteins (host protein sequences), the red intervening blocks depict the inteins. Letters within the proteins designate amino acids (single-letter code) that flank the splice sites. Residues printed in bold are predicted to be essential for the self-splicing mechanism.

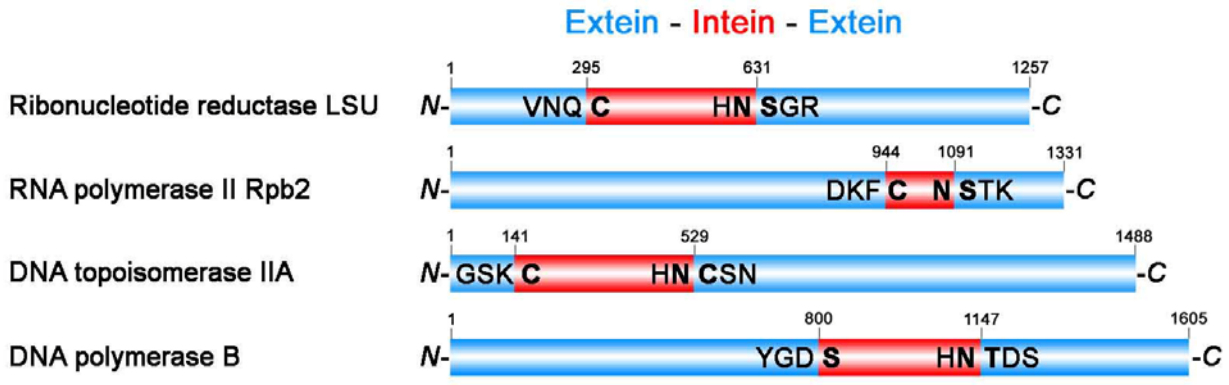


Table 2.4 Properties of the four CroV inteins.

The following categories represent a subset of the information collected on InBase, the NEB intein registry. 'DOD' is a member of the "LAGLIDADG" or dodecapeptide motif family of homing endonucleases. 'Prototype allele' lists the first intein identified at the identical extein insertion site in an extein homolog. 'Insert site' gives the name of the conserved extein site in which the intein is inserted.

	CroV inteins			
Intein name	**CroV Pol**	**CroV RIR1**	**CroV RPB2**	**CroV Top2**
CroV CDS	crov497	crov454	crov224	crov325
Extein name	DNA polymerase B	Ribonucleoside-diphosphate reductase, alpha subunit	DNA-dependent RNA polymerase II subunit 2	DNA Topoisomerase II beta
Allele Group	DNA polymerase I Motif C (pol-c insertion site)	Ribonucleotide-reductase (similar to Avi RIR1 BIL)	New	New
Prototype allele	Tli Pol-2	CroV RIR1	CroV RPB2	CroV Top2
Endonuclease motif	DOD (degenerate)	DOD (degenerate)	-	DOD
Location in extein	D799	Q294	F943	K140
Insert site	pol-c, Pol motif C	RIR1-h	RPB2-c	Top2-a
Intein size (aa)	347	336	146	387
N-terminal splice junction	VFNPIIKYGD/S	YNSICRYVNQ/C	ERPPIIGDKF/C	TGGMNGLGSK/C
C-terminal splice junction	HN/TDSIFSCYQY	HN/SGRRKGSFAF	GN/STKHGQKGTV	HN/CSNVWSKKFE

2.3.9 Repetitive DNA elements

About 5% of the CroV genome (excluding the unsequenced terminal regions) consists of repetitive DNA sequences (Figs. 2.5, 2.21). The most prevalent repetitive element is a 66 bp long repeat similar to the FNIP/IP22 repeat (Pfam entry PF05725), a member of the leucine-rich repeat (LRR) family. In total, there are more than 400 copies of this repeat in the CroV genome, and they are present in at least 28 CDSs (Fig. 2.5 and Table B.1 in Appendix B). Although the IP22 motif is degenerate, its widespread occurrence in the genome posed a major obstacle during contig assembly and gap-closing (see section 2.2.3 and Appendix A). IP22 is named after the first two conserved amino acids of this 22-amino acid-long motif [204]. A consensus sequence logo created from 258 IP22 units is shown in Fig. 2.22A. The CroV genome was found to contain two major and several minor versions of IP22 repeats; the predominant version 1 has a conserved tryptophane (W) at the beginning of the motif, whereas the less abundant version 2 starts with a leucine (L) (Fig. 2.22C). Both versions contain a highly conserved proline (P) at the second position. In accordance with the naming of IP22, I termed the CroV versions WP22 and LP22, respectively. Additional differences between WP22 and LP22 are found at position 7 (WP22: threonine, LP22: histidine), position 17 (WP22: proline, LP22: lysine), and in a single aa insertion at position 19 of the LP22 version (Fig. 2.22C). The distribution of these two versions in the genome was found to be non-random; the majority (>70%) of LP22 repeats is located within the first 32 kbp of the CroV genome, from which the WP22 repeat is completely absent (Fig. 2.21). In contrast, the WP22 repeat dominates the rest of the genome. Furthermore, the orientation of these repeats exhibits a clear pattern: all WP22/LP22 repeats located before position 470,000 are encoded by the reverse strand, whereas all WP22/LP22 repeats located after position 490,000 are encoded by the forward strand (Fig. 2.21). The biological reason for the observed distal orientation of the WP22/LP22 repeats is not known, but it is possible that they might be involved in replication or packaging of the viral genome. The IP22 repeat has been found in eight Mimivirus genes and more than 200 *Dictyostelium discoideum* genes [204], and is also present in the related dictyostelid *Polysphondylium pallidum* as well as in the phaeophyte *Ectocarpus siliculosus* [205]. Whereas LRRs are known to mediate protein-protein interactions in a variety of proteins with diverse functions [206] and IP22 repeats have been proposed to be involved in calmodulin binding and cell motility [204, 207], the specific role of these repeats in CroV is unknown.

Figure 2.21 Dot plot analysis of the CroV genome to show repeated sequences.

The linear CroV genome is plotted on both axes and matching sequences are marked by a dot at their cross position. The diagonal line represents the perfect match of the genome against itself. Any other structures are the result of repeated sequences in the genome. When the repeat sequence is in the same orientation as the source sequence, it is coloured in black; inverted repeats are coloured in red. Blue and green arrows below the dot plot indicate the orientation of LP22 and WP22 repeats, respectively. Numbers on axes refer to CroV genome positions.

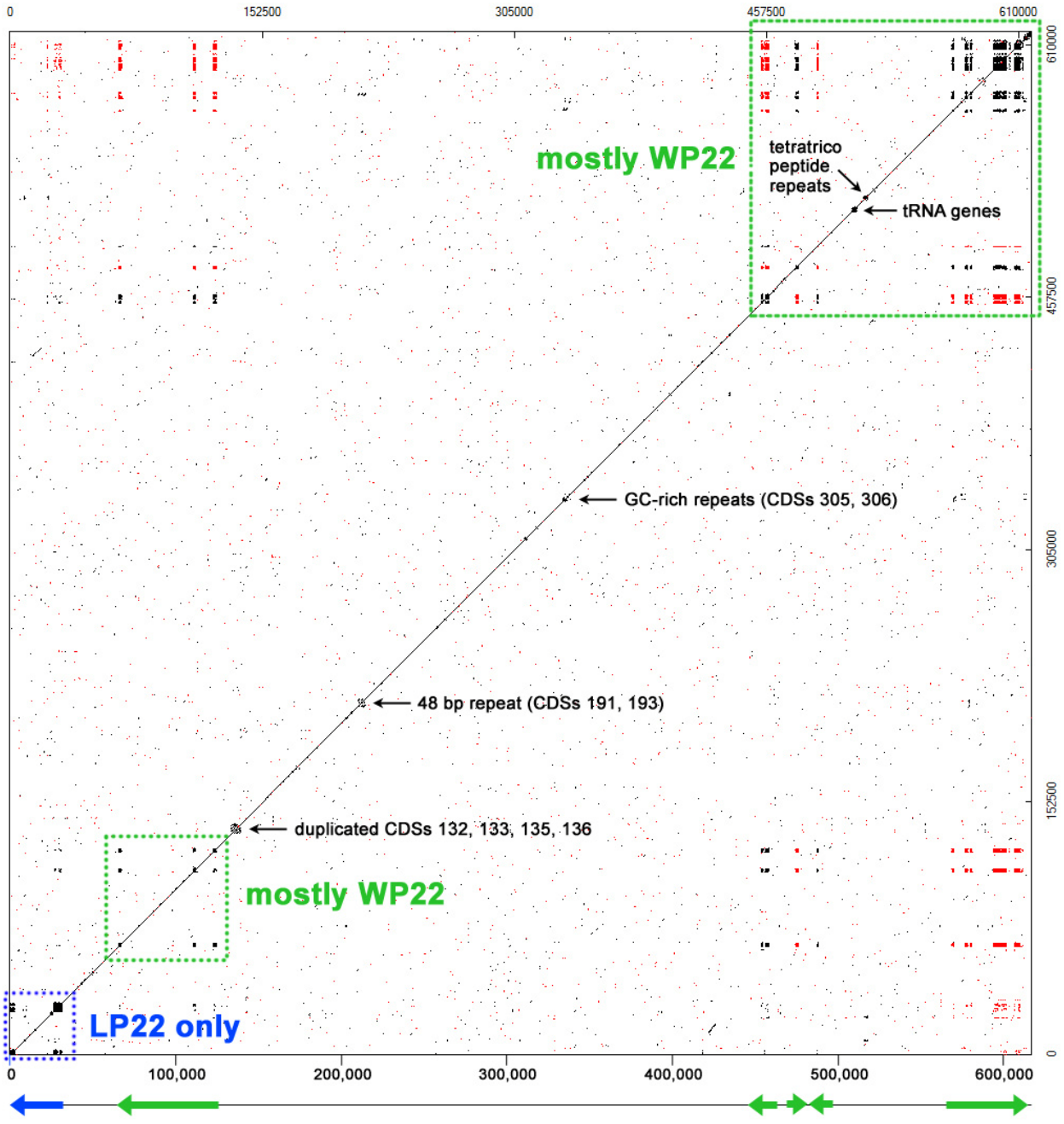
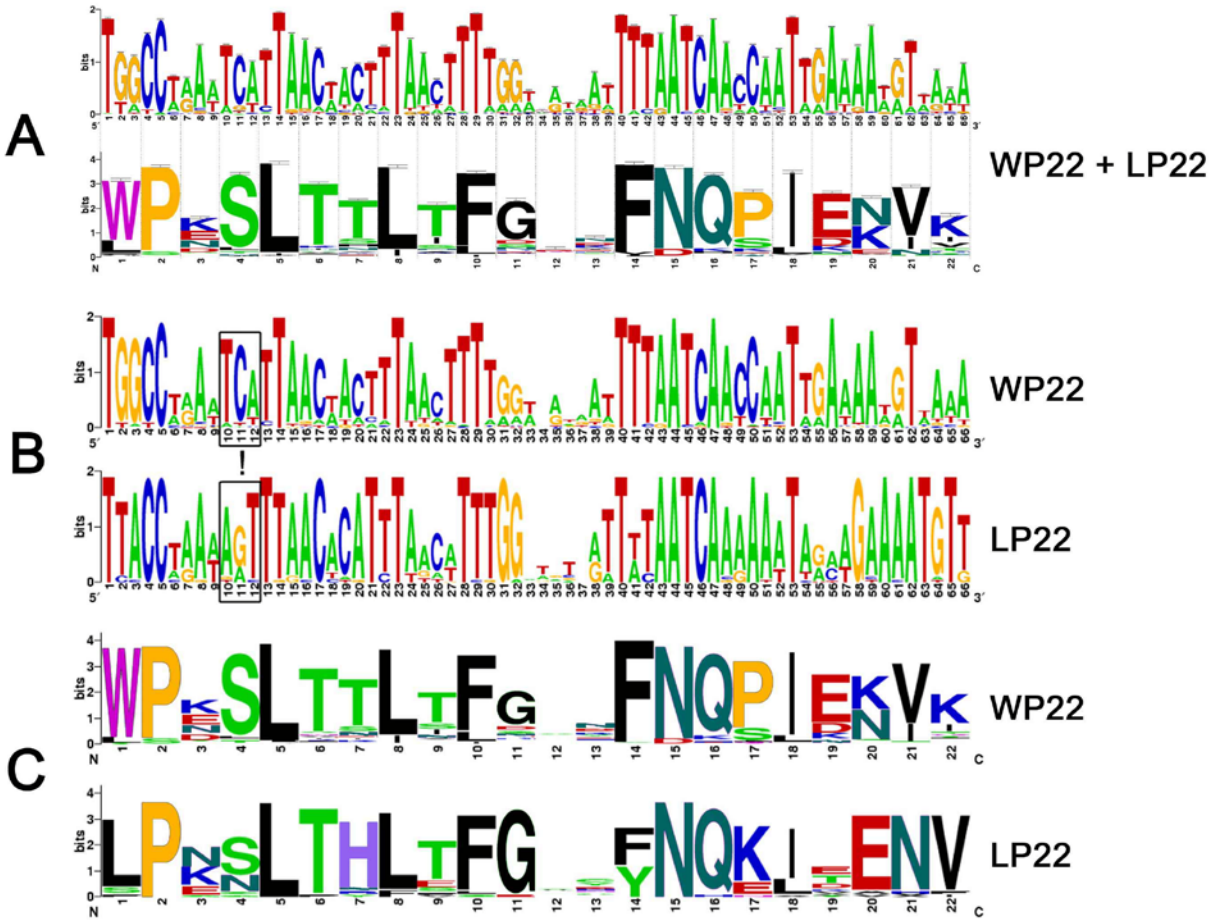


Figure 2.22 Sequence logos of FNIP/IP22 repeats in CroV.

(**A**) DNA and protein consensus sequence of 258 FNIP/IP22 repeats (all versions). (**B**) DNA logos created from alignments of 181 copies of the WP22 version, and 42 copies of the LP22 version, respectively. (**C**) Protein logos of the DNA logos from (B). All sequence logos were created using WebLogo[11] [154].



Despite considerable degeneracy of the WP22 motif (Fig. 2.22, amino acids 12 and 13), some residues are highly conserved at the protein level. For instance, proline is encoded by the triplet "CCN" ("N" being any nucleotide) and the conserved proline at position 2 of the WP22 repeat is encoded by highly conserved cytidines at the first two codon positions, but the third codon position is not conserved, as this has no influence on aa specificity (Fig. 2.22B,C). An even more striking example is the serine at position 4, which is encoded by the triplet "TCA" in the WP22 version, but by the unrelated "AGT" codon in the LP22 version (Fig. 2.22B, boxed residues). Furthermore, most of the WP22/LP22 repeats are found in-frame within a CDS. This suggests that these repeats are functionally conserved at the protein level.

[11] http://weblogo.berkeley.edu/logo.cgi

The WP22/LP22-containing proteins are likely to be the result of lineage-specific gene family expansion, where duplicated genes can acquire new functions [50].

Other forms of DNA repeats in the CroV genome are locally restricted and significantly less abundant than the WP22/LP22 repeat (Fig. 2.21). A 48 bp-long tandem repeat found in CDSs crov191 and crov193 translates into a 16 aa-long highly charged peptide with the consensus sequence "EKTKQEVEKTKQIEFI". Further repetitive elements include ankryin repeats (e.g. in crov429, crov430, crov440), GC-rich repeats in crov305 and crov306, and TPR domains in various CDSs such as crov467-471 (Fig. 2.21). Ankyrin repeats mediate protein-protein interactions and are found in almost all NCLDV genomes [50]; in Mimivirus, the 66 ankyrin-containing genes represent the largest paralogous family [56]. The helical TPR motif is composed of 34 loosely conserved amino acids and TPR domains are found in a variety of cellular proteins, where they mediate specific protein interactions [187] (see section 2.3.7).

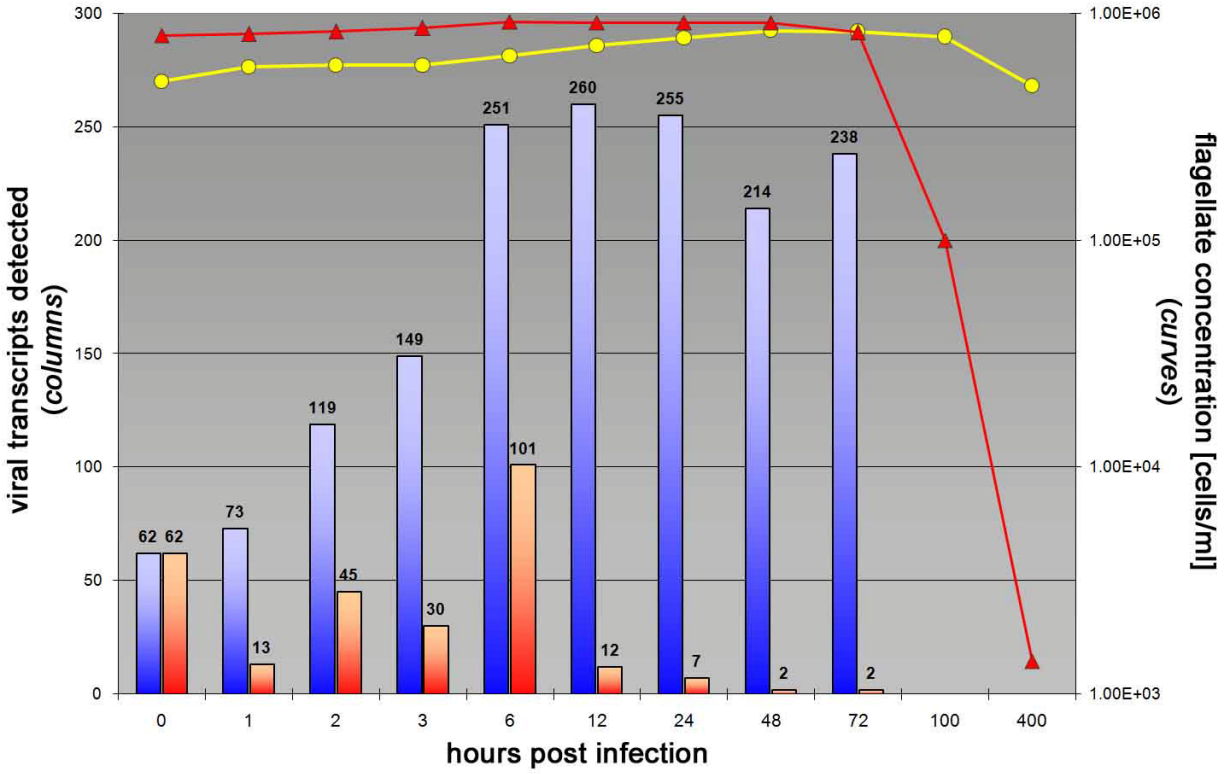
2.3.10 Microarray analysis

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. Viral transcripts were detected in infected *C. roenbergensis* cells by fluorescently labelling mRNA isolated at different time points during the infection cycle, which lasts 12-18 hours in *C. roenbergensis* strain E4-10. Labelled transcripts were then hybridized to glass slides spotted with oligonucleotide probes for 438 of the 544 predicted CroV genes, because the microarray was designed prior to completion of the genome sequence. A total of 274 genes (63%) had detectable levels of expression, 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. 2.5, Table B.1). About half of the predicted genes and 63% of the genes tested were expressed during infection. This percentage is comparable to the observed expression of 65% of viral genes during infection of the marine phytoplankter *Emiliania huxleyi* by EhV-86 [10]. However, recent gene expression studies in PBCV-1 and Mimivirus validated transcription for nearly all of their predicted genes [69, 208]. It seems therefore likely that the microarray data underestimates the true extent of transcriptional activity in CroV (see section 3.3.6 for further discussion and comparison with proteome data). All of the previously mentioned translation-related genes in CroV, as well as most of the "virus-atypical" genes were found to be expressed (Table B.1), 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 recognizing some general trends of

Figure 2.23 Temporal CroV gene expression profile during infection.

The x-axis shows the time points post infection at which RNA was extracted from infected cultures for microarray analysis (0-72 h p.i.) and two additional time points at which flagellate concentrations of infected (red curve) and uninfected (yellow curve) cultures were determined (100 h p.i., 400 h p.i.). Blue columns indicate the total number of CroV genes that were found to be transcribed at a given time point; red columns indicate the number of new transcripts, i.e. transcribed genes for which no transcripts had been found at previous time points. It should be noted that although most genes stayed "on" for the rest of the experiment once their mRNA had been detected, this was not the case for all genes; hence, the sum of red columns leading up to a certain time point is not equivalent to the value of the blue column at that time point.



CroV gene expression. Fig. 2.23 shows the temporal transcription profile of CroV infection. Immediately after infection, transcripts of 62 viral genes were detected. It is possible that some of these very early transcripts are already packaged in the virion, as similar findings have been reported from Mimivirus [26]. The number of newly transcribed genes decreased from 0 h p.i. to 1 h p.i., then increased slightly at 2 h p.i.. At 2 h p.i. and 3 h p.i., most of the transcription and replication enzymes were found to be expressed (Table B.1), suggesting that viral genome replication was initiated around 3 h p.i.. At 6 h p.i., there was a significant increase in transcriptional activity and capsid and core genes were then seen expressed for the first time. This suggests that the synthesis of capsid structures started as early as 3-6 h p.i.. From 12 h p.i. to 72 h p.i., only 21 genes were transcribed that had not been detected at earlier time points.

This is in agreement with electron microscopy observations of ultra thin-sectioned cells, suggesting that the CroV infection cycle lasts 12-18 hours (see chapter 4). The finding that infected cultures in the microarray experiment did not decline before 72 h p.i. can be explained by the mixed infection profile of this experiment. A relatively low MOI had to be used for inoculation in order to avoid instant cell lysis accompanied by a lack of CroV production, an effect that was observed at higher MOIs and may be similar to a mechanism known as "lysis from without" in other virus-host systems [209, 210]. Therefore, not all host cells were immediately infected at 0 h p.i.. Based on the time points at which transcripts were first detected, it was possible to distinguish between an early and a late stage of CroV gene expression. The early stage lasts from 0 h p.i. to 3 h p.i. and comprises 150 genes (Fig. 2.23, Table B.1). The majority of DNA replication and transcription genes belong to the early class. The late stage is characterized by genes that were first detected in the microarray at 6 h p.i. or later. The 124 genes in this class include 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, e.g. by subdividing the early stage into immediate early (0-1 h p.i.) and late early (2-3 h p.i.) stages, as hinted at by the temporal expression profile in Fig. 2.23.

2.3.11 Promoter analysis

The intergenic regions have an average size of 71 bp with a standard deviation of 64 bp. I examined the 100 nt region upstream of the predicted start codons for possible promoter motifs using MEME software [153]. A perfectly conserved "AAAAATTGA" motif, flanked by AT-rich sequences, was found to precede 127 CroV CDSs (23%) (Fig. 2.24A). The MEME E-value for this motif is 10^{-170}. Allowing one mismatch per sequence at the less strongly conserved positions one to six of the AAAAATTGA motif increases the number of positive CDSs to 191 (35%). The majority of CDSs that display this motif in their immediate upstream region belong to the "early" temporal category (pie chart in Fig. 2.24A). I therefore classified this motif as an early gene promoter in CroV. These results are in agreement with findings from Mimivirus, where a nearly identical motif (AAAATTGA) is present in the promoter region of 45% of Mimivirus genes, acting as an early promoter element [69, 70]. In contrast to Mimivirus, however, where the motif is found preferentially in the -50 to -110 region, the early promoter motif in CroV displays a much narrower distribution, with a peak at position -40 relative to the predicted start codon (Fig. 2.24B). The search for a possible late promoter motif was initiated with a representative set of six CDSs, all predicted to encode capsid components (see section 2.2.4). Five of these six genes exhibit the conserved tetramer "TCTA", flanked by AT-rich regions on either side,

Figure 2.24 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.

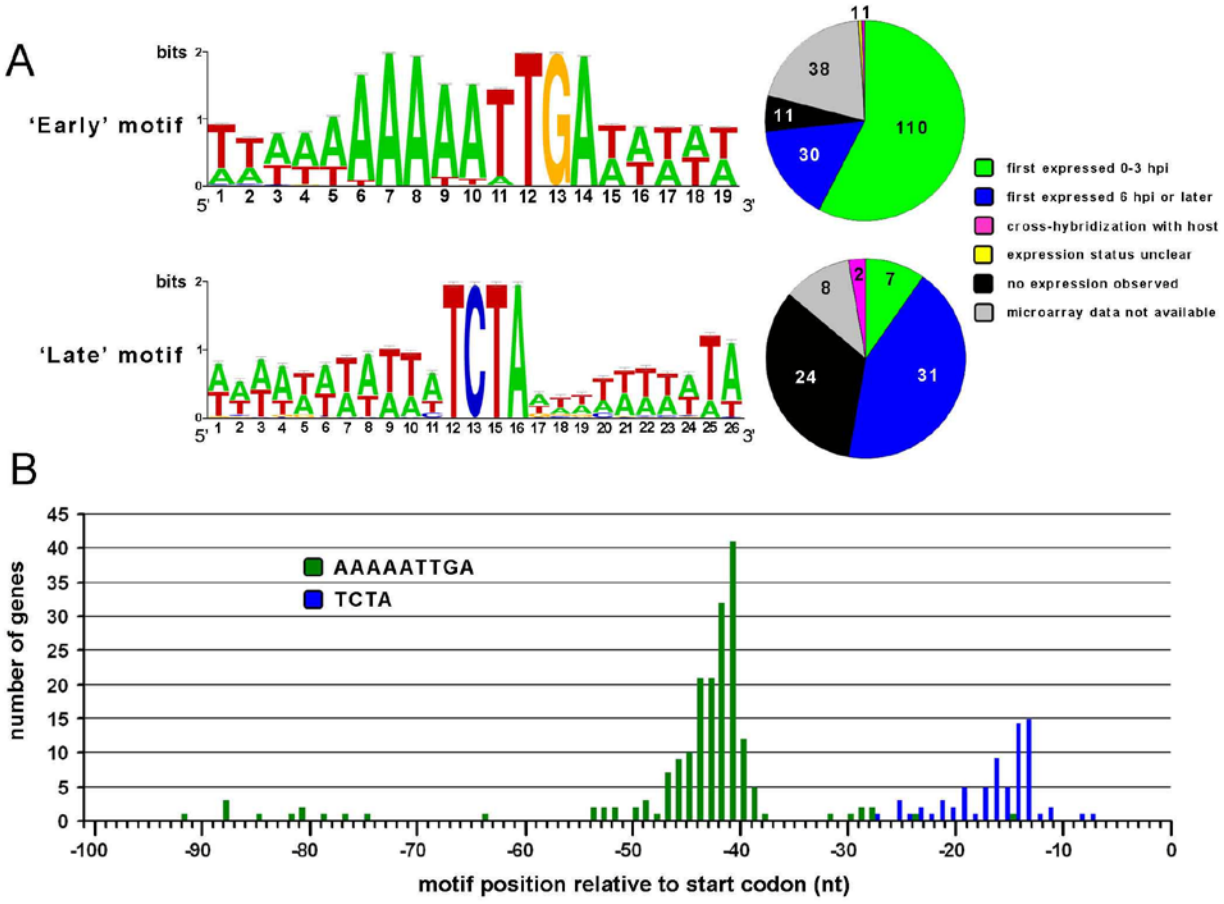
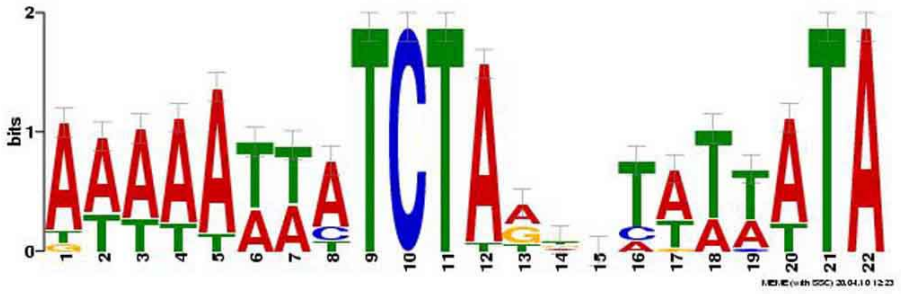


Figure 2.25 MEME sequence logo of the late promoter motif.



in their -11 to -20 region (Fig. 2.24). Based on this profile, I expanded the search to all CroV CDSs and found 72 CDSs to be 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 $5x10^{-4}$, Fig. 2.25). As shown in Fig. 2.24A, most CDSs with the TCTA promoter motif were first found to be expressed at 6 h p.i. or later, supporting the conclusion that this sequence motif represents a promoter element for genes transcribed during the late stage of CroV infection. The CroV late promoter motif is unrelated to the putative late promoter motif identified in Mimivirus [69].

2.3.12 A 38 kbp genomic fragment involved in carbohydrate metabolism

Upon examination of the CroV promoter distribution, I noticed that neither early nor late promoter motifs were associated with CDSs located between the genomic positions 264,800 and 302,500 (Fig. 2.5). Of these 34 CDSs (crov242 - crov275), 14 are most similar to bacterial proteins (Table 2.5, BLASTP E-value $<10^{-5}$) and seven of them are predicted to function in carbohydrate metabolism (Table B.1). Among them are enzymes for the biosynthesis of 3-deoxy-D-*manno*-octulosonate (KDO) (Fig. 2.26). In Gram-negative bacteria, KDO is an essential component of the lipopolysaccharide layer, linking lipid A to polysaccharides [211]. KDO is also found in the green alga *Chlorella* and in the cell wall of higher plants ([212] and references therein). Biosynthesis of KDO 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. 2.26A). Using conserved motif searches and multiple sequence alignments, crov265 was found to encode a predicted bifunctional KDOPase/API, and the N-terminal domain of crov267 contains a KDOPS homologue. 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 (Figs. 2.26B, 2.27-2.30). Whether the cytidylyltransferase in crov266 is a functional CMP-NeuAcS and accordingly involved in sialic acid activation, as suggested by phylogenetic analysis (Fig. 2.30), or rather a structurally related KDO-activating CKS, remains to be tested. The remaining CroV CDSs with functional annotation in this region are predicted glycosyl transferases and other sugar-modifying enzymes (Table B.1). The presence of these genes implies a role in viral glycoprotein biosynthesis and suggests that the virion surface is 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 display no similarity to Mimivirus genes, the 38 kbp 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 (Table 2.5, Figs. 2.27, 2.28), it is tempting to speculate that they may have been acquired from a bacterium, given that CroV frequently encounters phagocytosed bacteria inside the host cytoplasm. 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 towards the ends of the linear viral chromosomes [213]. However, given that the GC content of the 38 kbp 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 homologues (e.g. KDOPS and KDOPase, Figs. 2.28, 2.29), alternative scenarios for the origin of this region cannot be ruled out.

Figure 2.26 The predicted 3-deoxy-D-manno-octulosonate (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. Genomic coordinates are given in bp. DMKMT, demethylmenaquinone methyltransferase; TDP-DHMT, dTDP-6-deoxy-L-hexose 3-O-methyltransferase; CMP-NeuAcS, N-acylneuraminate cytidylyltransferase.

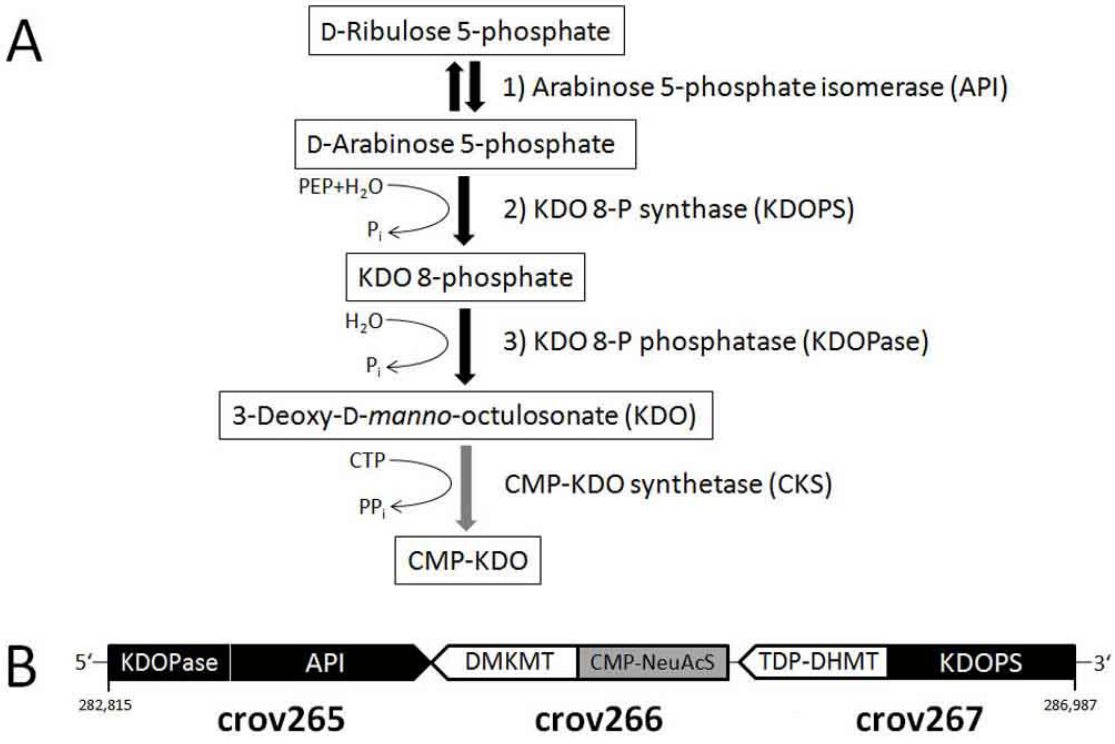


Table 2.5 Top BLASTP results for the 34 CDSs in the 38 kbp genomic fragment.

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	aa identity	Alignment length (aa)
crov242	UDP-glucose 6-dehydrogenase [*Bacteroides sp*. D4]	ZP_04557566	$9x10^{-23}$	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	$7x10^{-6}$	28%	164
crov245	-	-	-	-	-
crov246	alpha-2,3-sialyltransferase [*Campylobacter coli* RM2228]	ZP_00368088	$6x10^{-6}$	32%	143
crov247	glycosyl transferase family 2 [*Cyanothece sp*. PCC 7424]	YP_002381075	$3x10^{-7}$	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	$1x10^{-11}$	26%	298
crov251	chromosomal replication initiator protein DnaA [*Leptospira biflexa* serovar Patoc strain 'Patoc 1 (Paris)']	YP_001837424	3.7	31%	73
crov252	unknown protein [Sphingomonas sp. S88]	AAC44076	$3x10^{-9}$	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	$7x10^{-12}$	26%	235
crov255	hypothetical protein [Strongylocentrotus purpuratus]	XP_794970	2.9	34%	100
crov256	conserved hypothetical protein [*Bacteroides finegoldii* DSM 17565]	ZP_05416596	$4x10^{-12}$	30%	201
crov257	-	-	-	-	-
crov258	-	-	-	-	-
crov259	hypothetical protein NAEGRDRAFT_78502 [*Naegleria gruberi*]	XP_002681081	$5x10^{-14}$	30%	249
crov260	tetratricopeptide TPR_2 [*Arthrospira platensis* str. Paraca]	ZP_06380292	$6x10^{-35}$	34%	287
crov261	hypothetical protein GSU3022 [*Geobacter*	NP_954064	$8x10^{-41}$	40%	217

CroV CDS	Top BLASTP hit	Accession number	E-value	aa identity	Alignment length (aa)
	sulfurreducens PCA]				
crov262	predicted protein [*Thalassiosira pseudonana* CCMP1335]	XP_002288935	0.001	23%	198
crov263	hypothetical protein NY2A_B094R [*Paramecium bursaria* Chlorella virus NY2A]	YP_001497290	$3x10^{-20}$	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	$1x10^{-29}$ $5x10^{-35}$	41% 31%	158 297
crov266	NT+CT: cytidylyltransferase domain protein [gamma proteobacterium HTCC5015]	ZP_05062599	$9x10^{-88}$	41%	415
crov267	NT: 2-dehydro-3-deoxyphosphooctonate aldolase, putative [*Ricinus communis*] CT: hypothetical protein Tery_3108 [*Trichodesmium erythraeum* IMS101]	XP_002522197 YP_722714	$5x10^{-79}$ $1x10^{-29}$	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	$5x10^{-21}$	32%	264
crov270	-	-	-	-	-
crov271	tetratricopeptide TPR_2 repeat protein [*Arthrospira platensis* str. Paraca]	ZP_06381863	$6x10^{-13}$	26%	316
crov272	capsular protein [*Haloquadratum walsbyi* DSM 16790]	YP_659198	$8x10^{-10}$	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

Figure 2.27 Phylogenetic analysis of arabinose-5-phosphate isomerases.

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

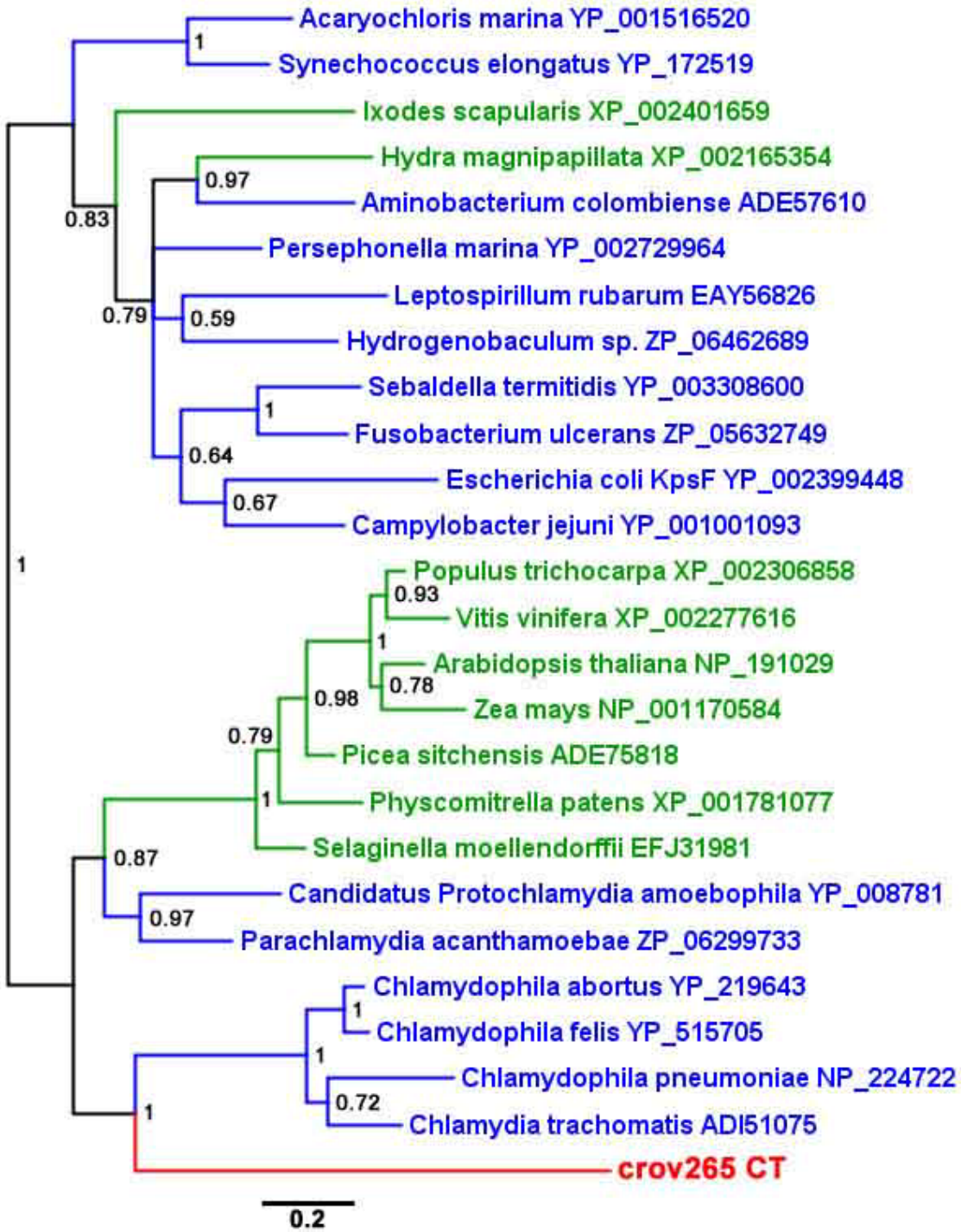


Figure 2.28 Phylogenetic analysis of KDO 8-P synthases.

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

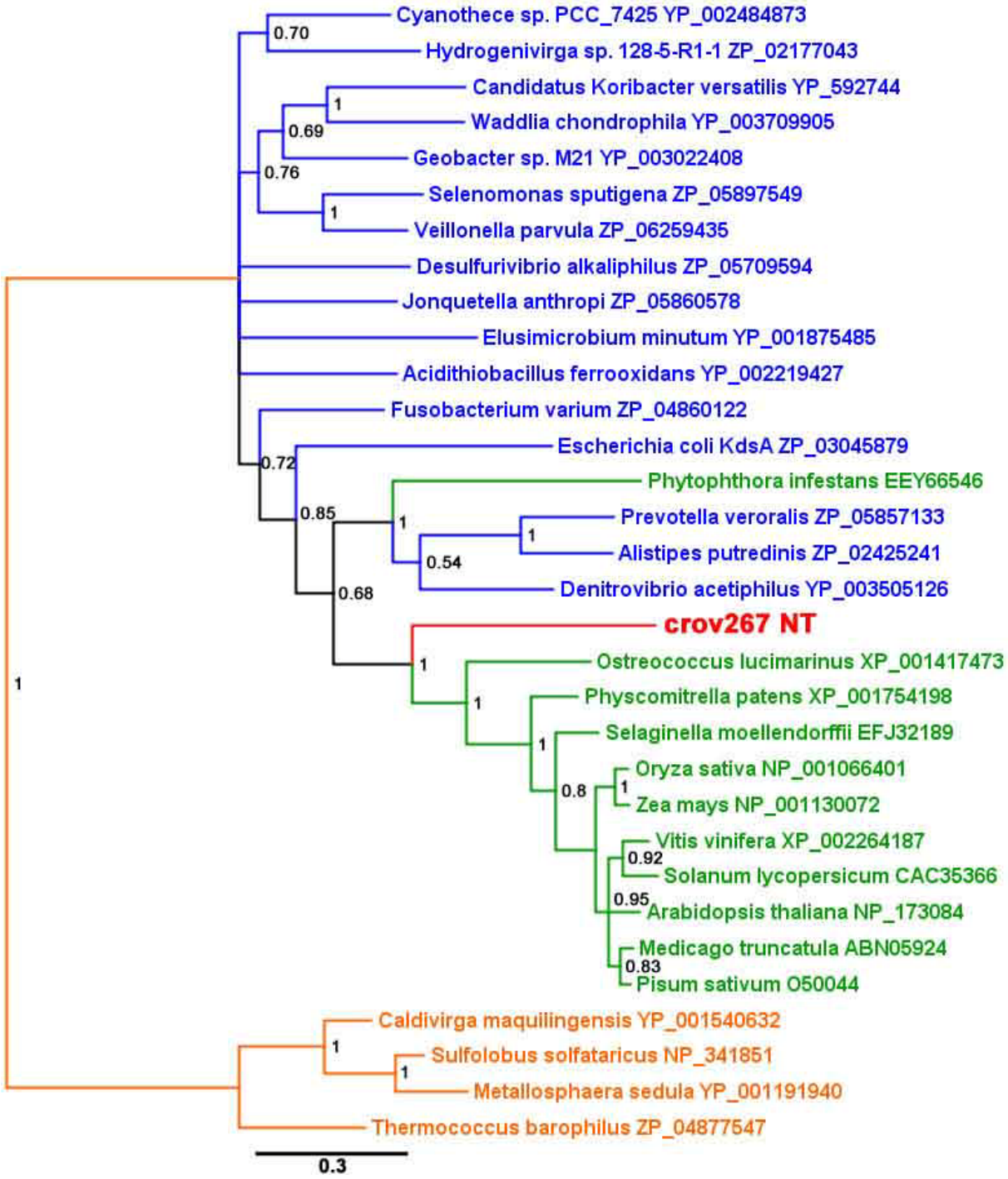


Figure 2.29 Phylogenetic analysis of KDO 8-P phosphatases.

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 an N-acetylneuraminate cytidylyltransferase domain. Sequences are coloured orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labelled with support values and GenBank accession numbers are given for each sequence.

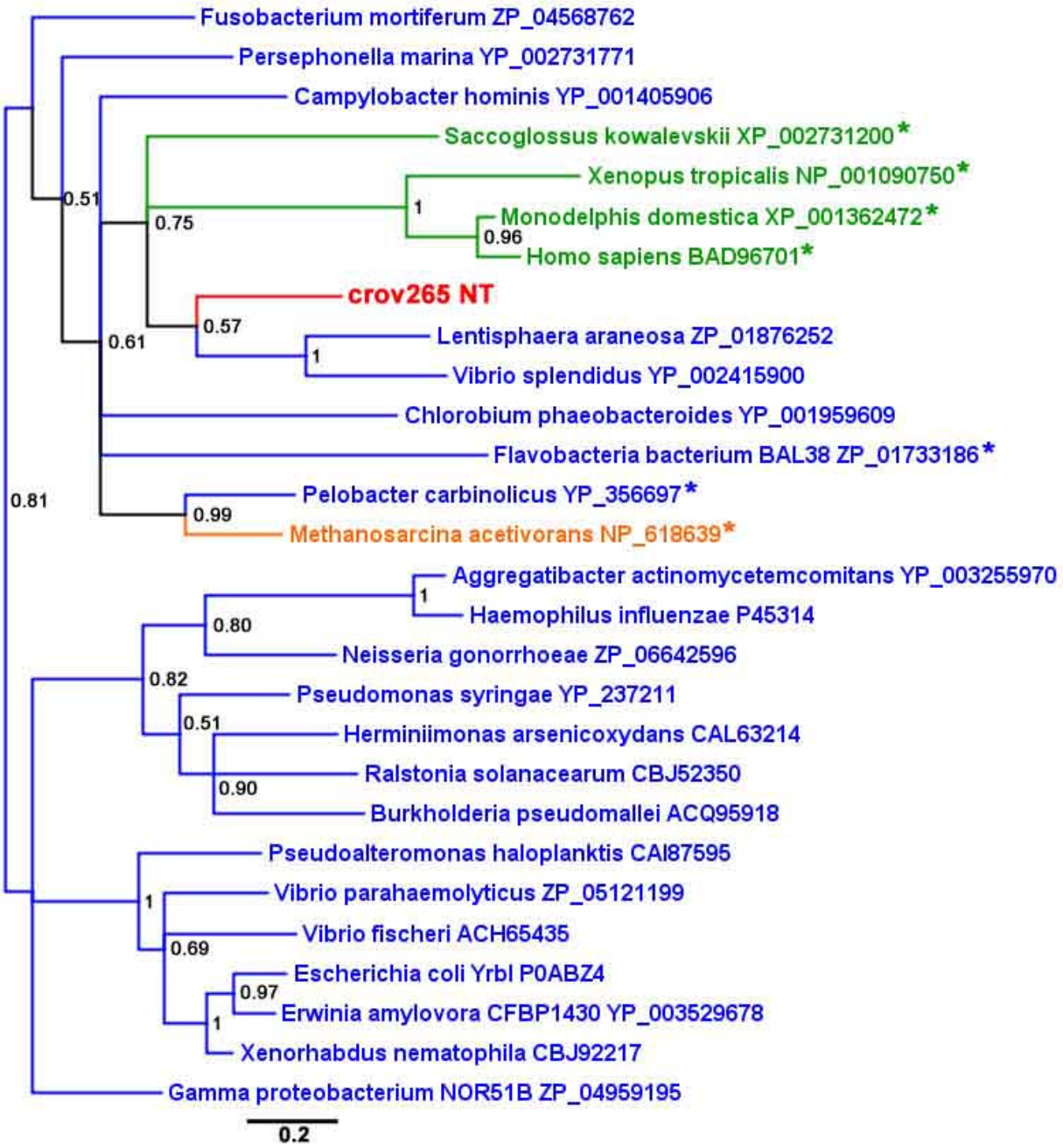
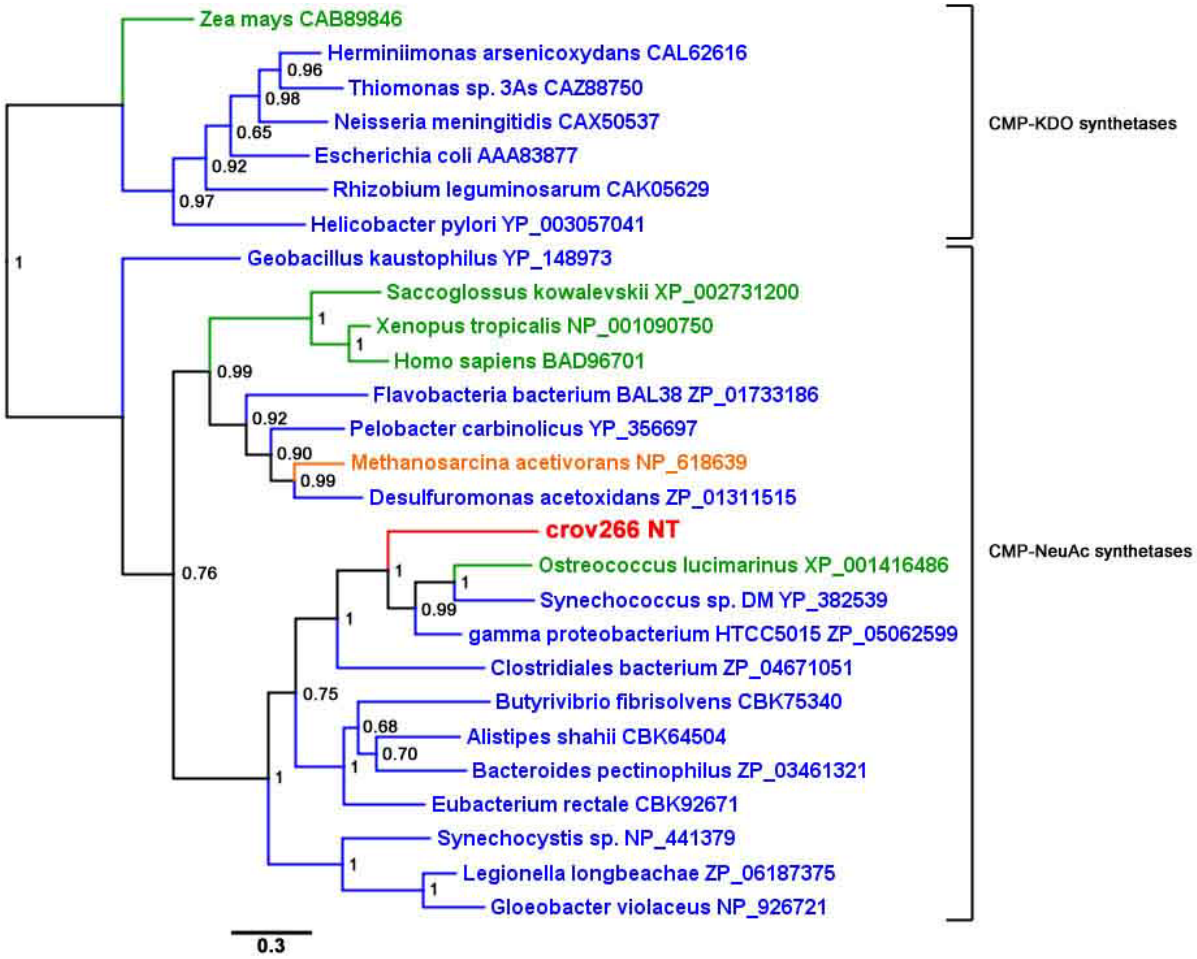


Figure 2.30 Phylogenetic analysis of two types of cytidylyltransferases.

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 coloured orange for archaea, blue for bacteria, and green for eukaryotes. Nodes are labelled with support values and GenBank accession numbers are given for each sequence.



2.3.13 Phylogenetic relationship to other viruses

Based on the presence and phylogenetic analysis of a set of core genes (Fig. 2.31), CroV is a new addition to the presumably monophyletic group of NCLDVs [26, 49, 50], which includes the families *Ascoviridae, Asfarviridae*, *Iridoviridae*, *Mimiviridae*, *Phycodnaviridae*, *Poxviridae*, and the newly discovered Marseillevirus [39]. In a recent study by Yutin and coworkers, genes encoded by NCLDVs were categorized into groups that presumably evolved from a common ancestor and subsequently diversified in the various NCLDV families [51]. Using this dataset of NucleoCytoplasmic Virus Orthologous Genes (NCVOGs), at least 172 CroV CDSs belong to an existing NCVOG (Table B.1). Thirty-two percent of CroV CDSs are significantly similar to a Mimivirus gene (any Mimivirus hit with a BLASTP E-value < 10^{-5}) and 22 CroV CDSs have their only detectable GenBank homologue in Mimivirus (as of September 2010). CroV therefore represents the closest known relative to Mimivirus, despite large differences in genome (730 kbp vs. 1,181 kbp) and capsid size (280 nm vs. 500 nm). The CroV-Mimivirus relationship is further corroborated by phylogenetic analysis of the DNA polymerase PolB, a commonly used marker gene to infer phylogenetic relationships among NCLDVs. Bayesian Inference analysis of conserved regions 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. 2.32). These three algal viruses, for which only PolB and major capsid protein (MCP) sequences were available, also possess very large DNA genomes (485 kbp, 510 kbp, and 560 kbp, respectively) and are proposed members of the family *Phycodnaviridae*, although a taxonomic revision of this tentative assignment has been suggested [98]. Similarly, when the MCP is used to reconstruct the NCLDV phylogeny (Fig. 2.33), these five viruses form a monophyletic group that also includes *Heterosigma akashiwo* virus, another giant 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 large genome, probably encoding more than 150 proteins.

Figure 2.31 NCLDV core genes found in CroV.

Shown are NCLDV core genes of groups I-III present in CroV and selected members of the NCLDV. Viral hallmark genes are bolded. *Mimivirus L451 is a putative RuvC-like Holliday Junction Resolvase (HJR) homologue (18). **OtV-1_053 was identified as a putative RuvC-like HJR homologue.

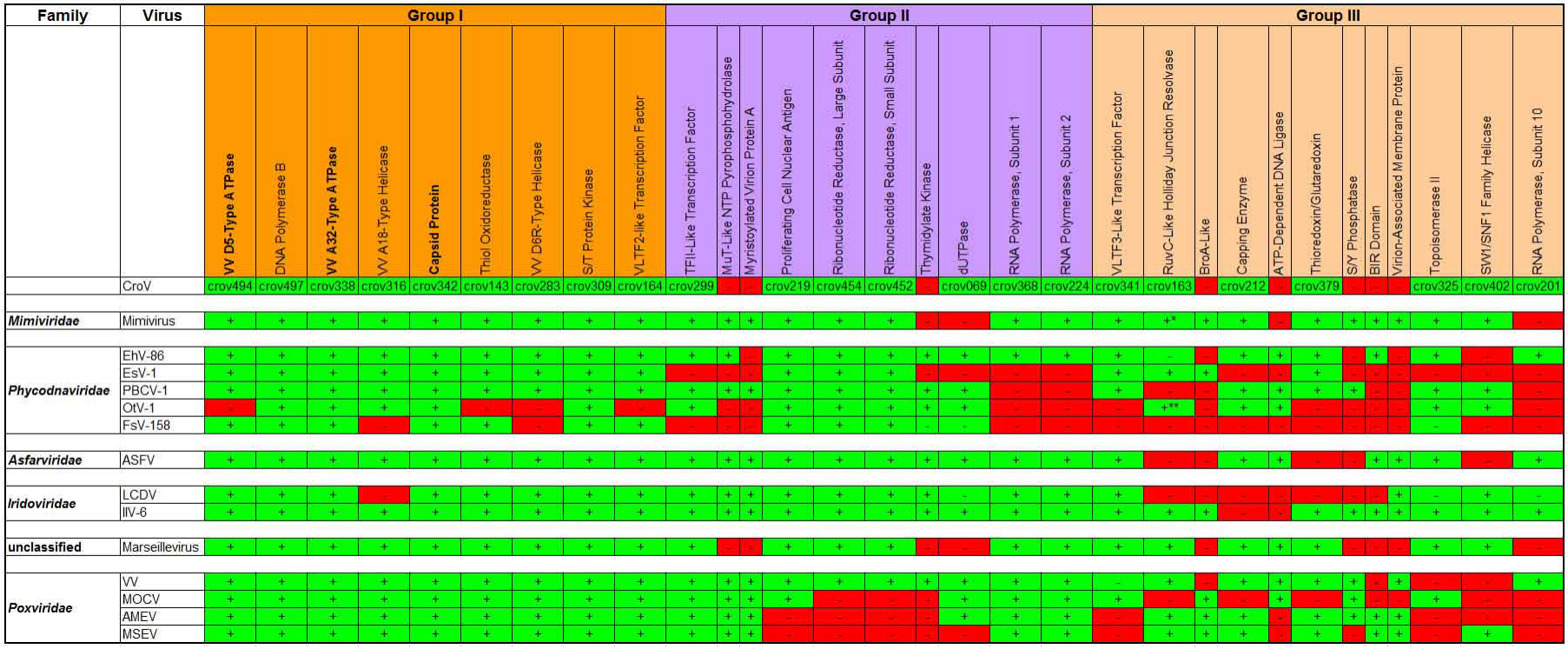


Figure 2.32 Phylogenetic reconstruction of NCLDV members based on DNA polymerase B.
The unrooted Bayesian Inference tree was generated from a 263 aa alignment of conserved regions of DNA polymerase B. Intein insertions were removed prior to alignment. Nodes are labelled with posterior probabilities and ML bootstrap values (500 replicates). Abbreviations and accession numbers (GenBank unless stated otherwise) are as follows: ACMV, *Acanthamoeba castellanii* mamavirus, from [51]; AMV, *Amsacta moorei* entomopoxvirus, NP_064832; APMV, *Acanthamoeba polyphaga* mimivirus, YP_003986825; ASFV, African swine fever virus, NP_042783; ATCV-1, *Acanthocystis turfacea* virus 1, YP_001427279; CeV-01, *Chrysochromulina ericina* virus 01, ABU23716; CIV, Chilo iridiscent virus, NP_149500; CroV, *Cafeteria roenbergensis* virus, ADO67531; DpAV4, *Diadromus pulchellus* ascovirus 4a, CAC19127; EhV-86, *Emiliania 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, *Heterosigma akashiwo* virus 01, BAE06251; HcDNAV, *Heterocapsa circularisquama* DNA virus, DDBJ accession number AB522601; HvAV3, *Heliothis virescens* ascovirus 3e, YP_001110854; IIV-3, Invertebrate iridiscent virus 3, YP_654692; ISNKV, 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, *Paramecium bursarium* chlorella virus 1, NP_048532; PoV-01, *Pyramimonas orientalis* virus 01, ABU23717; PpV-01, *Phaeocystis pouchetti* virus 01, ABU23718; TnAV2, *Trichoplusia ni* ascovirus 2c, YP_803224; VV, Vaccinia virus, AAA98419.

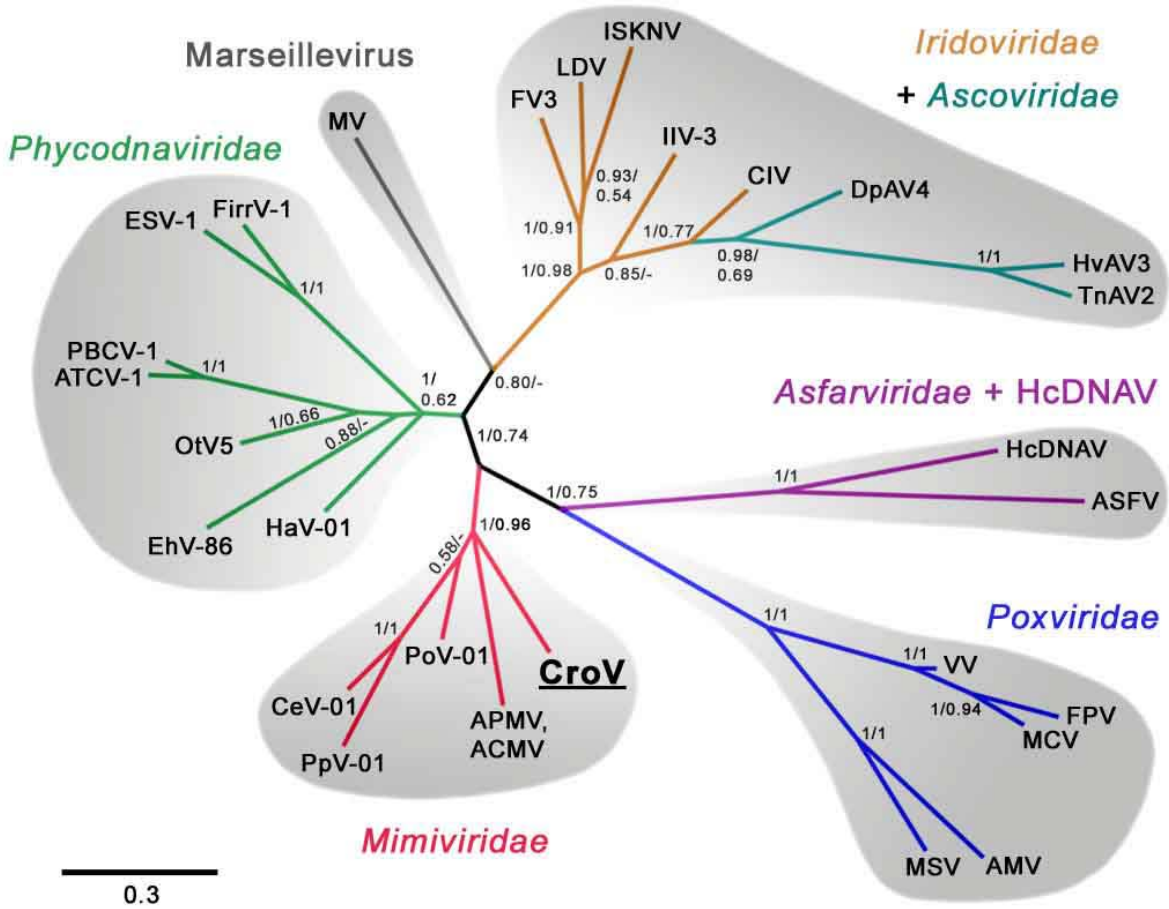
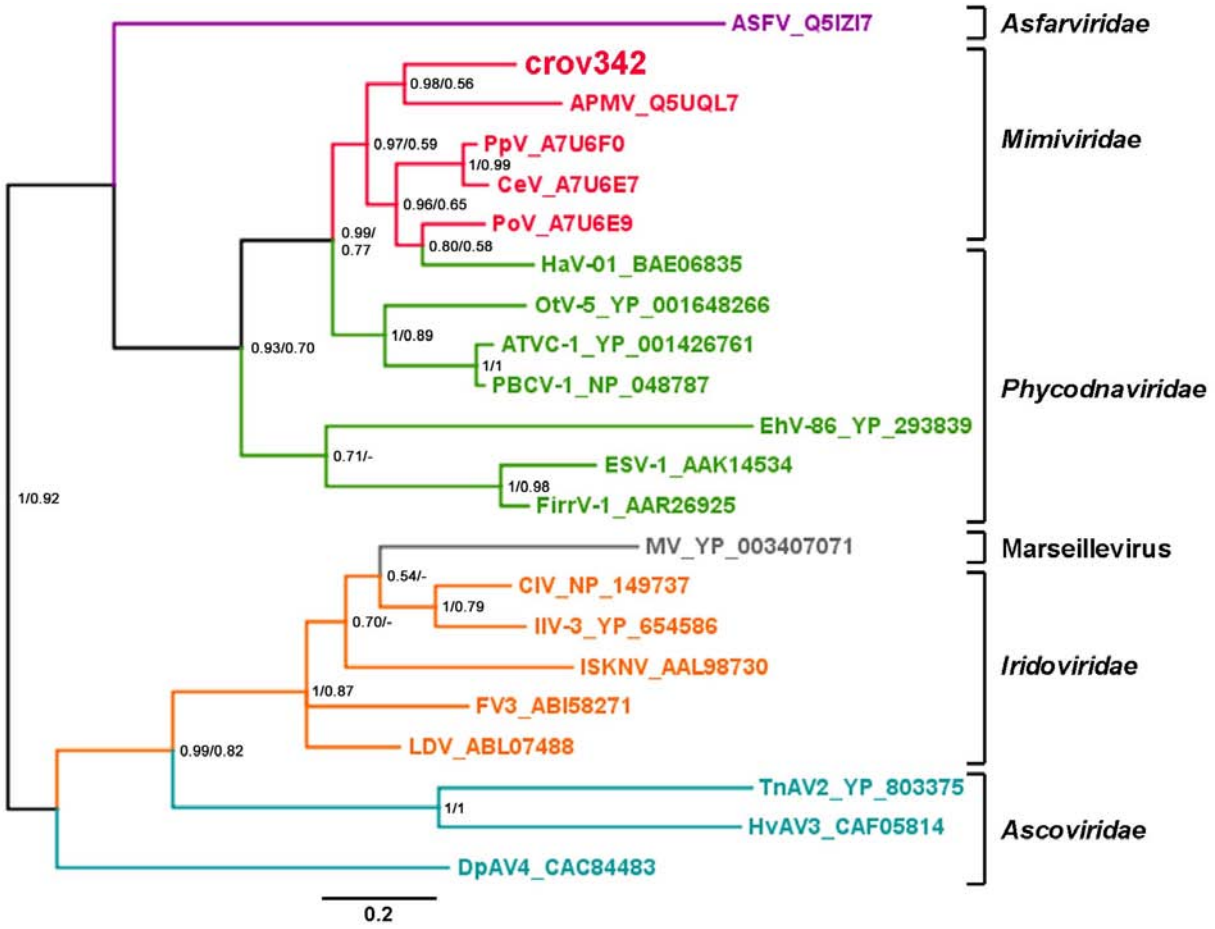


Figure 2.33 Phylogenetic analysis of the NCLDV major capsid protein.

The unrooted Bayesian Inference tree is based on a 169 aa alignment. Colour coding and abbreviations are the same as in Fig. 2.32. Not included in the tree are the poxviruses, as their capsid proteins are too divergent from those of other NCLDV families. Nodes are labelled with posterior probabilities and ML bootstrap values (500 replicates). GenBank accession numbers are given for each sequence.



2.4 Conclusions

This chapter presents the first genome 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 [56, 213, 214]. Most studies have focused on Mimivirus, since it represents the most extreme case of a giant virus and is the largest dataset available. The majority of Mimivirus genes have no cellular homologues and are presumably very ancient [165], up to one third of its genes arose through gene and genome duplication [56], and less than 15% of Mimivirus genes may have been horizontally transferred from eukaryotes and bacteria [84]. This general picture of giant virus genome evolution seems to be confirmed by my initial analysis of the CroV genome. Gene duplication and lineage-specific expansion of the FNIP-like 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 crov027-031 (contain LP22 repeats), crov420-422 (unknown function), and some of the tRNA genes. A potential case of large-scale HGT from a bacterium is represented by the 38 kbp genomic segment that differs in coding content and promoter regions from the rest of the viral genome (see section 2.3.12). The remaining CDSs with cellular homologues are more difficult to categorize, since genes can be transferred from cells to viruses and vice 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 continues to close 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 non-living adds to the ongoing debate about the puzzling evolutionary history of giant viruses [48, 130, 214].

3 The Virion Proteome of CroV

3.1 Synopsis

The virion is the physical entity of a virus that protects the genetic material and delivers it to a suitable host cell. As such, the proteins present in the virion can give important clues about the extracellular, penetration, and early intracellular stages of the virus replication cycle. In this chapter, the protein composition of the giant *Cafeteria roenbergensis* virus (CroV) particle is assessed. Liquid chromatography-tandem mass spectrometry (LC-MS/MS) was performed on purified CroV virions and resulted in the identification of 129 virus-encoded proteins. In addition to structural components, CroV packages many non-structural proteins, including 17 transcription proteins, 6 DNA repair enzymes, 5 redox-active proteins, and 4 proteases. Seventy-eight proteins identified in this study had no known function. A comparison of CroV and Mimivirus revealed that 34 proteins are present in both virions. The finding that giant viruses infecting hosts of different eukaryotic supergroups share a conserved set of virion proteins suggests that the last common ancestor of these viruses must have already possessed a similarly complex particle composition. These results also imply that the early events of CroV infection may resemble those of Mimivirus and the poxviruses, where viral transcription starts in the host cytoplasm while the viral genome is still enclosed in the virion core.

3.2 Materials and methods

3.2.1 Virus purification

The culturing conditions of *Cafeteria roenbergensis* strain E4-10 have been described in section 2.2.1. Exponentially growing cultures were infected with CroV, and cell concentrations were monitored by staining with Lugol's Acid Iodine and counting by microscopy using a hemocytometer. CroV virions were purified from ~10 l of culture lysate by centrifugation for 1 h at 10,500 x g in a Sorvall RC-5C centrifuge (GSA rotor, 4°C) to pellet bacteria and cell debris. CroV was concentrated from the supernatant by tangential flow filtration (TFF) using a Vivaflow 200, 0.2 µm PES unit (Sartorius Mechatronics, Canada) to a final volume of ~30 ml. Fourteen millilitres of concentrate was poured into Ultra-clear™ ultracentrifuge tubes (Beckman, Canada), then 0.4 ml of 50% (w/v) Optiprep™ solution (Sigma, Canada) was carefully pipetted to the bottom of the tubes, and the tubes were centrifuged for 1 h in a Sorvall RC80 ultracentrifuge (SW40 rotor, 20°C). A centrifugal force of 150,000 x g was necessary to concentrate CroV in the high buoyancy sea water medium. The concentrated layer of virus and contaminating

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 mMTris-HCl, pH 7.6, 4°C. The dialyzed suspension was loaded in 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 mMTris-HCl, pH 7.6), which had been linearized at room temperature overnight. The gradient was centrifuged for 1.5 h at 110,000 x g (SW40 rotor, 20°C), after which the CroV-containing fraction 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 further by centrifugation for 1 h at 20,000 x g, 20°C, in 1.5 ml microfuge tubes in an Eppendorf 5810 tabletop centrifuge and the pellets were resuspended in a final volume of 20 µl.

3.2.2 Protein sample fractionation and tryptic digestion[12]

The 20 µl sample containing ~$3x10^{10}$ viruses was mixed with $2.5x10^{-3}$ U/µl benzonase nuclease (Novagen, Germany) and 1% SDS for 1 h at room temperature. Proteins were resolved on a 10% SDS-polyacrylamide gel which was stained with blue-silver colloidal Coomassie [215]. The gel lane was cut into 10 bands followed by in-gel reduction (10 mM dithiothreitol for 45 min at 56°C), alkylation (55 mM iodoacetamide for 30 min at 37°C) and trypsinization (12.5 ng/µl of trypsin overnight at 37°C).

3.2.3 Liquid chromatography-tandem mass spectrometry

The samples were desalted with a STop-And-Go Extraction (STAGE) tip [216] and analyzed by liquid chromatography-tandem mass spectrometry using a linear trapping quadrupole-Orbitrap (LTQ-OrbitrapXL, Thermo Fisher Scientific, Germany). This was coupled on-line to a 1100 Series nanoflow high performance liquid chromatography system (HPLC; Agilent Technologies) that uses a nanospray ionization source (Proxeon, Denmark) and a holding column packed in-house with ReproSil-Pur C18-AQ, 3 µm, 120 A (Dr. Maisch GmBH) into 20 cm long and 50 µm inner diameter [217]. A trap column containing AQUA C18, 5 µm, 200 A (Phenomenex) was also used. Buffer A (0.5% acetic acid) and buffer B (0.5% acetic acid, 80% acetonitrile) were run in gradients of 10% B to 32% B over 57 min, 32% B to 40% B over 5 min, 40% B to 100% B over 2 min, 100% B for 2 min, and finally 0% B for another 10 min to recondition the column [217]. The LTQ-Orbitrap was set to acquire a full-range scan at 60,000 resolution from 350 to

[12] Protein sample fractionation, digestion, and LC-MS/MS analysis have been conducted by Dr. I. Kelly in the laboratory of Dr. L. Foster, University of British Columbia.

1500 Th in the Orbitrap and to simultaneously fragment the top five peptide ions in each cycle in the LTQ.

In addition, peptides were analyzed on a Q-TOF (quadrupol - time of flight) mass spectrometer (Accurate Mass 6520 Q-TOF, Agilent), which was on-line coupled to an Agilent 1200 Series nanoflow HPLC using a Chip Qube electrospray ionization Interface (Agilent). The Agilent HPLC consisted of a 1200 Series binary capillary pump with a vacuum degasser for sample loading and a nano-pump with a vacuum degasser, autosampler and thermostat for gradient separation. A high capacity HPLC chip with 160 nl enrichment column and a 0.075 mm x 150 mm analytical column containing Zorbax 300SBC-18 5 µm stationary phase (Agilent) was used in the Chip Qube. The thermostat temperature was set at 6°C. The autosampler was set for a draw speed of 5 µl/min and an ejection speed of 40 µl/min. Each sample was automatically loaded on the enrichment column at flow rate 4 µl/min of buffer A (0.1% formic acid, 3% acetonitrile) and at 4 µl injection flush volume. Gradients of 110.2 min were run with the nano-pump at 0.3 µl/min from 0% buffer B (0.1% formic acid, 90% acetonitrile) to 5% B over 2 min, 5% B to 35% B over 78 min, 35% to 45% B over 10 min, 45% to 95% over 0.1 min, 95% B for 20 min, and finally 95% to 3% B over 0.1 min to recondition the column for the next analysis. The Chip Qube was set at a degas heater temperature of 325°C and a degas flow rate of 5 l/min. The ion optics were set at the following values: VCap 1860 V, Fragmentor 175 V, Skimmer 1 65 V, Octopole RF Peak 750 V, MS1 Centroid Data Abs Threshold 20, MS2 Centroid Data Abs Threshold 5. The Q-TOF was set to acquire a full-range scan on the TOF detector from 300 to 2000 Th at a scan rate of 6 for MS and from 50 to 3200 Th at a scan rate of 3 for MS/MS (isolation width 4 amu).

Collision energy was calculated automatically depending on the charge state of the parent ions using a collision energy slope of 3.6 and collision energy offset of 2.4. A maximum of 6 parent ions (absolute precursor threshold 1000, relative precursor threshold 0.01, charge state 2+ and higher) were selected for MS/MS after each MS scan and were excluded for 0.25 min. The data were processed using Spectrum Mill Rev A.03.03.084 SR4 with a parent mass tolerance of 10 ppm, a fragment mass tolerance of 1000 ppm, fixed modification carbamidomethylation, low parent 300.0 Th, high parent 6000 Th, maximum sequence tag length 1, group scans 15 s, parent tolerance 1.4, similarity merging 1, minimum signal/noise ratio PMF 15, minimum MS signal/noise 25.

3.2.4 Data processing

Fragment spectra were searched against a custom protein database consisting of 543 predicted CroV protein-coding sequences (CDSs), 1353 CroV unidentified reading frames (URFs; these are open reading frames [ORFs] that are unlikely to encode proteins because they are shorter than 150 nt or overlap with a larger ORF, see section 2.2.4), 20 virophage CDSs, 53 virophage URFs, and 229 ORFs from a bacteriophage called phage307 (see Appendix E). The search was conducted using Mascot (v2.2, Matrix Science) with the following parameters: trypsin specificity allowing up to one missed cleavage; pI range 3.0-10.0; cysteine carbamidomethylation as a fixed modification; methionine oxidation and deamidation of asparagine and glutamine as variable modifications; ESI-trap fragmentation characteristics; 10 ppm mass tolerance for precursor ion masses; 0.8 Da tolerance for fragment ion masses. Proteins were considered identified when at least two unique peptides of eight or more amino acids were found with Mascot Ions Scores >25, resulting in an estimated false discovery rate of less than 0.5% based on reversed database searching. The relative abundances of protein components were estimated by dividing the number of peptides identified in MS/MS experiments by the number of expected peptides obtained by *in silico* trypsination. Transmembrane helices in viral proteins were predicted using the Hidden Markov Model implemented in the TMHMM Server v. 2.0[13] [218].

[13] http://www.cbs.dtu.dk/services/TMHMM

3.3 Results and discussion

3.3.1 Identification of CroV virion proteins

Viruses are obligate intracellular parasites and must establish a productive infection in their hosts in order to propagate themselves. The virion composition is a critical factor in this process, as it determines host specificity and mediates penetration and entry of virus particles into the host cell. Virion proteins are easily amenable to proteomic methods if the particles can be obtained in pure form, and several virion proteomes have already been described [68, 219-226]. CroV virions were purified by Optiprep^{TM} density gradient centrifugation, denatured and freed of nucleic acids. Analysis by LC-MS/MS identified 128 CroV CDSs, 1 CroV URF, 1 Mavirus CDS, and 2 phage307 CDSs. Table 3.1 lists all viral proteins identified in this study. The single CroV URF identified by LC-MS/MS was only 47 aa long and therefore below the length cutoff of 50 aa used for gene finding in the CroV genome (see section 2.2.4). This URF was subsequently upgraded to CDS crov299a. A search of the GenBank non-redundant database was performed to identify potential contaminants and returned significant hits for trypsin (the enzyme used for protein digestion), benzonase from *Serratia marcescens* (the enzyme used for nucleic acid degradation), keratin (a common contaminant introduced during sample handling), and actin from *Bos taurus*. No other eukaryotic proteins were found, suggesting that the purified CroV sample was free of cellular debris from lysed flagellates. However, with no sequence information available about the host genome, it was not possible to conduct a targeted search for host protein contamination. The absence of bacterial proteins indicates that the CroV purification protocol was highly efficient in removing the mixed bacterial assembly that was present in enriched seawater cultures of *C. roenbergensis*. A relative abundance (RA) index was calculated for each identified protein as an approximate measure of protein stoichiometry, based on the number of obtained mass spectra and the number of expected spectra based on protein sequence; RA values do not reflect absolute protein quantities.

Genes coding for proteins packaged in the CroV virion are distributed relatively uniformly across the CroV genome, as is revealed when these genes are ordered against their gene numbers (Fig. 3.1). Therefore, virion proteins are not preferentially encoded by any particular region of the genome. However, a large gap in the middle of the plot indicates that the 42 proteins encoded by CDSs crov241 to crov282 are absent from the virion proteome. These CDSs coincide with a 38 kbp genomic region that was found to differ from the rest of the CroV genome in terms of promoter regions and coding content and is presumably of bacterial origin (see

section 2.3.12). The genes in this region code for enzymes involved in carbohydrate metabolism, including a complete biosynthetic pathway of 3-deoxy-D-*manno*-octulosonate (also known as KDO), which is an essential component of the cell walls in plants and bacteria [227]. None of the proteins from this region were found to be packaged in the virion. The terminal regions of the CroV genome also deviate slightly from the central part (Fig. 3.1). The slope of the first 10 virion proteins is steeper than the rest, indicating that fewer proteins from the 5'-end of the assembled genome are packaged. Similarly, the 3'-terminal region of the assembled genome is only sparsely represented in the virion proteome, as only one of the proteins encoded by crov499-crov544 was identified by LC-MS/MS analysis. These observations can potentially be ascribed to the transitory nature of these genomic regions, being located between the highly repetitive terminal regions (for which no sequence information is available) and the central protein-coding part of the genome. With 129 virus-encoded proteins packaged, CroV possesses one of the most complex virus particles known. Of the 129 CroV proteins, 82 are of unknown function; this represents 64% of the virion proteins and is comparable to the 57% of proteins found in the Mimivirus virion proteome that were of unknown function [68]. Among the CroV proteins with functional annotations, five major categories can be distinguished: virion structure, transcription, DNA repair, protein modification, and oxidative pathways (Table 3.2).

Figure 3.1 Approximate genomic distribution of genes encoding virion proteins

The data points represent CroV proteins identified by LC-MS/MS, plotted in increasing order of their gene numbers.

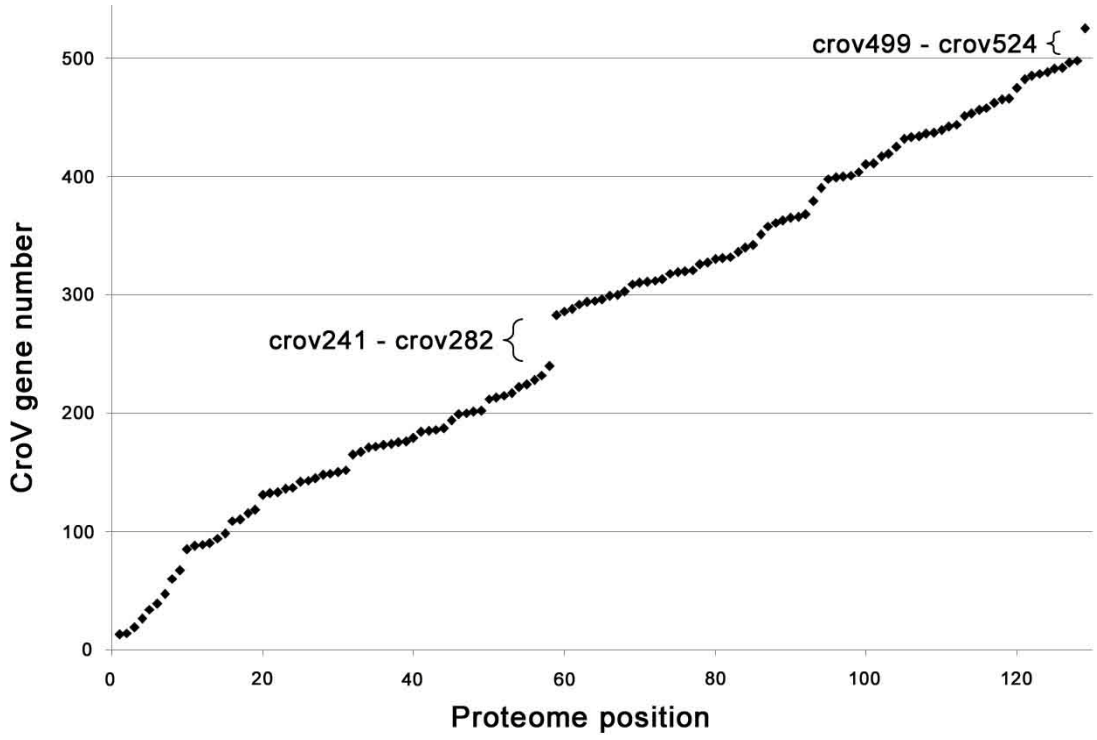


Table 3.1 Viral proteins identified in this study through LC-MS/MS analysis.

'Expression' indicates the time point at which the respective gene was first found to be expressed in a microarray experiment (section 2.3.10). The column 'Mimi hom.' lists homologous proteins in Mimivirus. In case of several homologues, only the best matching sequence is shown, except for cases where the protein is encoded by a single CDS in CroV, but two or more adjacent CDSs in Mimivirus. Mimivirus proteins in bold are confirmed virion components. Abbreviations: h p.i., hours post infection; N/A, data not available; N/C, expression status not clear; N/D, transcript not detected; host, microarray probes for this transcript cross-hybridized with a host cell transcript; TMH, number of predicted transmembrane helices.

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
CroV												
Virion structure	crov342	major capsid protein	57.24	1027	1426	2453	51	48.10	late	6	-	**L425**
Virion structure	crov398	capsid protein 2	53.58	30	35	65	73	0.89	late	6	-	**L425**
Virion structure	crov321	capsid protein 3	76.63	2	4	6	75	0.08	late	N/D	-	R439
Virion structure	crov176	capsid protein 4	58.77	2	0	2	72	0.03	-	N/D	-	R441
Virion structure	crov332	major core protein	67.33	749	1189	1938	89	21.78	late	6	-	**L410**
Virion structure	crov148	phage tail collar domain protein	26.22	100	114	214	20	10.70	late	6	-	-
Virion structure	crov187	predicted structural protein	14.42	84	135	219	19	11.53	late	12	-	-
Virion structure	crov240	predicted structural protein	8.37	31	63	94	2	47.00	late	N/A	1	-

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
Virion structure	crov318	predicted structural protein	9.33	52	91	143	14	10.21	-	N/A	-	**L485**
Virion structure	crov320	predicted structural protein	20.00	123	181	304	11	27.64	late	6	3	**-**
Transcription	crov368	DNA-directed RNA polymerase II, subunit Rpb1	169.46	149	147	296	229	1.29	early	2	-	**R501**
Transcription	crov224	DNA directed RNA polymerase, subunit Rpb2 precursor	152.40	165	116	281	218	1.29	early	2	-	**L244**
Transcription	crov482	DNA-dependent RNA polymerase II, subunit Rpb3/Rpb11	37.93	24	36	60	43	1.40	early	0	-	**R470**
Transcription	crov439	DNA-directed RNA polymerase, subunit Rpb5	22.36	22	18	40	32	1.25	early	6	-	**L235**
Transcription	crov491	DNA-dependent RNA polymerase II, subunit Rpb6	17.71	5	11	16	23	0.70	early	2	-	R209
Transcription	crov437	DNA-directed RNA polymerase II, subunit Rpb7	27.15	9	10	19	25	0.76	early, late	2	-	**L376**
Transcription	crov492	DNA-directed RNA polymerase II, subunit Rpb9	21.27	11	28	39	34	1.15	early	1	-	**L208**
Transcription	crov201	DNA-directed RNA polymerase, subunit Rpb10	9.07	12	9	21	11	1.91	-	N/A	-	**-**

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
Transcription	crov109	VV D6-like very early transcription factor, small subunit	132.21	85	34	119	176	0.68	late	6	-	**L377**
Transcription	crov292	VV A7-like very early transcription factor, large subunit	217.13	36	39	75	300	0.25	-	1	-	**R326/ R327**
Transcription	crov299	transcription elongation factor TFIIS	18.48	3	2	5	26	0.19	early	2	-	R339
Transcription	crov283	VV D11-like transcription factor	90.87	119	82	201	130	1.55	late	3	-	**R350**
Transcription	crov131	VV D11-like transcription factor	66.25	4	3	7	94	0.07	-	N/C	-	**R563**
Transcription	crov212	mRNA capping enzyme	119.40	100	61	161	158	1.02	-	3	-	**R382**
Transcription	crov286	poly(A) polymerase catalytic subunit	53.72	37	44	81	73	1.11	-	host	-	**R341**
Transcription	crov118	RNA helicase	193.75	7	3	10	260	0.04	-	N/D	-	**R366**
Transcription	crov152	DNA topoisomerase IB	62.14	13	11	24	89	0.27	-	N/C	-	**R194***
DNA repair	crov115	CPD class I photolyase	56.68	3	4	7	75	0.09	-	N/D	-	-
DNA repair	crov149	prokaryotic photolyase	58.23	2	0	2	74	0.03	-	N/D	-	R852/ R853 R855
DNA repair	crov200	exonuclease	41.72	12	14	26	60	0.43	-	N/D	-	**L533**
DNA repair	crov303	formamidopyrimidine-DNA glycosylase	33.30	3	0	3	52	0.06	-	6	-	L315
DNA repair	crov458	family X DNA polymerase	40.98	10	8	18	69	0.26	early	3	-	**L318**

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
DNA repair	crov462	NAD-dependent DNA ligase	74.84	4	4	8	125	0.06	-	2	-	R303
Redox pathway	crov013	contains N-terminal oxidoreductase domain	55.96	10	10	20	87	0.23	-	N/D	1	-
Redox pathway	crov143	Erv1/Alr family thiol oxidoreductase	21.36	9	6	15	24	0.63	-	6	-	**R596**
Redox pathway	crov444	Erv1/Alr family thiol oxidoreductase	19.29	14	16	30	26	1.15	-	N/A	1	R368
Redox pathway	crov172	protein disulfide isomerase	17.49	33	32	65	19	3.42	late	N/D	1	**R443**
Redox pathway	crov379	protein disulfide isomerase	13.93	9	15	24	19	1.26	late	N/D	-	**R443**
Protein modification	crov067	metalloendopeptidase	72.68	16	13	29	95	0.31	-	6	-	R519
Protein modification	crov137	insulinase-like metalloprotease	102.12	9	4	13	129	0.10	-	0	-	L233
Protein modification	crov184	cysteine protease	34.71	10	13	23	46	0.50	-	6	-	**R355**
Protein modification	crov485	cysteine protease	31.14	12	5	17	30	0.57	-	N/D	-	-
Protein modification	crov309	serine/threonine protein kinase	46.62	7	6	13	60	0.22	-	0	-	**R400**
Protein modification	crov475	family 2C serine/threonine phosphatase	30.50	14	14	28	42	0.67	-	6	-	**R306/ R307**
Protein modification	crov525	serine/threonine protein phosphatase	73.18	19	8	27	92	0.29	-	0	-	**-**

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
Methylation	crov442	FtsJ-like methyltransferase	52.42	14	24	38	64	0.59	early	24	-	**R383**
Amino acid transport	crov232	ABC-type amino acid transporter	56.94	12	42	54	45	1.20	early	0	3	-
Diverse	crov171	ADP-ribosyl glycohydrolase	39.91	4	2	6	39	0.15	early	N/D	-	L444
N/A	crov417	contains SET domain	18.12	2	6	8	25	0.32	-	N/D	-	-
N/A	crov319	contains ankyrin repeats	165.57	107	139	246	222	1.11	-	N/D	-	**L484**
N/A	crov313	phosphoesterase	36.92	2	3	5	49	0.10	-	2	-	**R398**
N/A	crov014	unknown	14.28	5	5	10	9	1.11	-	N/D	3	-
N/A	crov019	unknown	58.44	43	32	75	93	0.81	-	N/D	-	-
N/A	crov026	unknown	34.99	5	3	8	57	0.14	-	N/D	-	-
N/A	crov034	unknown	19.37	4	3	7	23	0.30	-	N/D	-	R871
N/A	crov039	unknown	17.75	84	72	156	22	7.09	late	N/D	-	-
N/A	crov047	unknown	106.45	6	5	11	52	0.21	late	6	-	-
N/A	crov060	unknown	28.49	3	4	7	31	0.23	-	48	1	-
N/A	crov085	unknown	11.20	2	0	2	13	0.15	late	N/A	1	-
N/A	crov088	unknown	80.46	15	15	30	90	0.33	late	72	1	-
N/A	crov089	unknown	49.32	2	0	2	44	0.05	-	6	1	-
N/A	crov090	unknown	45.66	31	24	55	47	1.17	late	N/D	1	-
N/A	crov094	unknown	21.80	58	125	183	27	6.78	late	6	-	-
N/A	crov098	unknown	9.04	2	11	13	16	0.81	late	N/A	-	-
N/A	crov110	unknown	36.21	65	69	134	47	2.85	-	2	-	-

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
N/A	crov132	unknown	58.07	19	0	19	65	0.29	late	0	-	-
N/A	crov133	unknown	55.72	33	41	74	67	1.10	-	N/C	-	-
N/A	crov136	unknown	63.94	29	48	77	74	1.04	late	6	-	-
N/A	crov142	unknown	210.28	21	18	39	134	0.29	-	6	-	-
N/A	crov145	unknown	29.24	35	38	73	27	2.70	late	N/D	-	-
N/A	crov150	unknown	94.17	3	0	3	126	0.02	-	3	-	-
N/A	crov165	unknown	21.79	6	7	13	32	0.41	-	N/D	-	-
N/A	crov167	unknown	14.09	12	25	37	15	2.47	late	N/D	1	-
N/A	crov173	unknown	32.14	21	33	54	39	1.38	-	N/D	-	-
N/A	crov174	unknown	46.85	153	193	346	64	5.41	late	6	-	-
N/A	crov175	unknown	28.07	174	178	352	40	8.80	late	N/D	1	-
N/A	crov179	unknown	13.18	11	9	20	22	0.91	-	N/D	-	-
N/A	crov185	unknown	98.74	75	24	99	129	0.77	-	6	-	**L454**
N/A	crov186	unknown	22.12	90	98	188	35	5.37	late	N/D	-	**R457**
N/A	crov194	unknown	22.14	4	15	19	25	0.76	late	12	2	**-**
N/A	crov199	unknown	81.58	71	76	147	97	1.52	-	6	-	**-**
N/A	crov202	unknown	12.16	27	21	48	17	2.82	-	N/D	1	**-**
N/A	crov213	unknown	16.81	4	9	13	22	0.59	late	N/D	-	**-**
N/A	crov215	unknown	20.73	20	12	32	25	1.28	-	6	-	**L492**
N/A	crov217	unknown	101.45	325	444	769	138	5.57	late	6	-	**-**
N/A	crov222	unknown	45.85	24	20	44	58	0.76	-	0	1	**-**
N/A	crov228	unknown	23.76	41	63	104	13	8.00	late	6	1	**R489**
N/A	crov288	unknown	21.82	11	19	30	25	1.20	-	6	-	-

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
N/A	crov294	unknown	11.68	7	33	40	13	3.08	late	N/D	2	-
N/A	crov295	unknown	20.33	3	5	8	29	0.28	-	3	-	-
N/A	crov296	unknown	23.19	25	17	42	33	1.27	late	N/D	-	-
N/A	crov299a (formerly URF718)	unknown	5.56	2	22	24	7	3.43	-	N/A	-	-
N/A	crov310	unknown	34.43	6	3	9	41	0.22	-	N/D	1	-
N/A	crov311	unknown	183.08	131	121	252	251	1.00	-	N/D	-	**R472**
N/A	crov312	unknown	39.44	34	39	73	25	2.92	late	6	-	-
N/A	crov326	unknown	75.90	5	0	5	77	0.06	-	N/D	-	R481
N/A	crov327	unknown	38.75	7	0	7	19	0.37	late	6	1	-
N/A	crov330	unknown	125.93	216	198	414	158	2.62	-	6	-	**R402 R403**
N/A	crov331	unknown	20.53	5	0	5	25	0.20	-	N/D	-	R409
N/A	crov336	unknown	51.50	41	65	106	66	1.61	late	6	-	**L417**
N/A	crov340	unknown	11.33	14	10	24	11	2.18	-	N/A	1	-
N/A	crov351	unknown	15.69	3	0	3	25	0.12	-	3	-	L202
N/A	crov358	unknown	39.45	3	9	12	61	0.20	late	0	-	-
N/A	crov361	unknown	20.69	11	23	34	21	1.62	late	6	2	-
N/A	crov363	unknown	39.99	28	27	55	42	1.31	late	host	1	-
N/A	crov365	unknown	14.38	17	19	36	6	6.00	late	6	3	-
N/A	crov366	unknown	8.56	3	7	10	12	0.83	late	N/A	1	-
N/A	crov390	unknown	119.10	2	3	5	55	0.09	-	6	-	R554

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
N/A	crov399	unknown	43.39	3	4	7	56	0.13	-	6	-	-
N/A	crov400	unknown	41.43	44	65	109	49	2.22	late	N/D	-	-
N/A	crov401	unknown	10.57	9	8	17	13	1.31	-	N/A	1	-
N/A	crov404	unknown	15.54	37	14	51	9	5.67	late	6	2	-
N/A	crov410	unknown	12.71	6	17	23	17	1.35	-	0	-	-
N/A	crov411	unknown	17.90	16	23	39	14	2.79	late	12	1	-
N/A	crov419	unknown	56.65	40	50	90	92	0.98	late	N/D	-	-
N/A	crov425	unknown	15.13	8	11	19	15	1.27	-	N/D	4	-
N/A	crov432	unknown	30.99	6	9	15	41	0.37	-	3	-	-
N/A	crov433	unknown	15.41	34	50	84	14	6.00	-	0	-	-
N/A	crov434	unknown	6.70	13	19	32	9	3.56	-	N/A	-	-
N/A	crov436	unknown	16.10	20	32	52	18	2.89	-	N/D	-	-
N/A	crov451	unknown	24.94	2	10	12	15	0.80	-	6	1	-
N/A	crov453	unknown	40.93	5	8	13	50	0.26	-	6	-	-
N/A	crov456	unknown	72.26	7	2	9	74	0.12	-	N/D	-	L371
N/A	crov465	unknown	13.09	3	2	5	16	0.31	-	N/D	-	-
N/A	crov466	unknown	23.42	4	2	6	30	0.20	-	N/D	1	-
N/A	crov487	unknown	12.04	22	14	36	16	2.25	-	N/D	-	-
N/A	crov488	unknown	16.41	4	5	9	17	0.53	-	N/D	-	-
N/A	crov496	unknown	25.14	5	8	13	15	0.87	-	N/D	1	-
N/A	crov498	unknown	173.25	82	81	163	220	0.74	-	0	-	-

Functional category	CDS	Putative function	Predicted mass (kDa)	Orbi-trap spectra	Q-TOF spec-tra	Total spec-tra	Expected peptides	Relative protein abundance	Gene promo-ter type	Expression (h p.i.)	TMH	Mimi hom.
Mavirus virophage												
Virion structure	CDS18	Major capsid protein	66.34	7	3	10	79	0.127	CroV late	N/A	-	-
Phage307												
Virion structure	ORF58 (331 aa)	Major capsid protein	37.02	17	13	30	52	0.58	N/A	N/A	-	-
N/A	ORF60 (135 aa)	unknown	14.04	7	8	15	7	2.14	N/A	N/A	-	-

Table 3.2 Functional categories represented by CroV virion proteins.

Proteins were binned by function based on matches to Pfam, Interpro, and NCBI COG and nr databases (see section 2.2.4).

Functional category	Number of proteins found in CroV virion
Transcription	17
Protein modification & turnover	7
DNA repair	6
Structure	6
Oxidative pathways	5
Others	6
Unknown	82

3.3.2 Structural proteins

All four predicted capsid proteins encoded by CroV were identified in this study, albeit with very different RA values (Table 3.1). The major capsid protein is the most abundant structural protein with 2453 mass spectra and an RA value of 48.10, followed by the major core protein with 1938 spectra and an RA value of 21.78. Another predicted structural component present in the virion is a phage tail collar domain protein (214 spectra, RA = 10.70). It is noteworthy that no peptides were found to match the predicted P9/A32-like DNA-pumping ATPase of the FtsK-HerA family (CROV338), which are molecular motors that use ATP to translocate viral genomes into prefabricated capsids [228]. The P9/A32 ATPase is a known virion component in internal membrane-containing dsDNA viruses such as bacteriophage PRD1 [229], but has not been detected in proteomic studies of vaccinia virus (VV) [225, 230-232] or EhV-86 [224], and was statistically only poorly supported in the Mimivirus proteome [59]. Therefore, the genome packaging ATPases in these viruses may be only transiently associated with the virion during particle assembly or may be present in too few copies to be detected by mass spectrometry.

Apart from these proteins, which are homologues of well-described structural proteins, it is likely that a complex virus particle such as the CroV virion consists of many additional proteinaceous components. To identify potential structural candidates among the 82 proteins of unknown function, I compared the RA values of predicted structural and non-structural proteins. Among the 41 annotated non-structural proteins, the predicted protein disulfide isomerase crov172 had the highest RA value (3.42), whereas the three most abundant structural proteins had RA values ranging from 10.70 to 48.10 (Table 3.1). Based on the observed difference, I considered any protein with an RA value of 10 or greater as a candidate for a structural virion component. This was the case for CDSs crov187, crov240, crov318, and crov320, which had RA values of 11.53, 47.00, 10.21, and 27.64, respectively. These four proteins may therefore represent major constituents of the CroV virion structure. Since two of them (crov240 and crov320) are predicted transmembrane (TM) proteins, they may interact with the internal lipid membranes (see section 5.3.4). A total of 33 proteins (26%) identified in this study are predicted to contain one or more transmembrane helices (TMHs, Table 3.1). This is far less than the 82% (23 of 28) of coccolithovirus EhV-86 virion proteins that are predicted to contain TMHs [224]. However, EhV-86 contains an outer membrane, which it acquires during budding from the host membrane [233], 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 TM proteins, as revealed by TMHMM analysis of the Mimivirus virion proteome [68]. The total number of predicted TM proteins encoded by CroV

is 69, and 33 of them are present in the virion. Therefore, proteins with predicted TMHs are overrepresented in the CroV virion (Fisher's exact test, p-value $2.7x10^{-6}$), suggesting that TM proteins are essential for the CroV virion, e.g. for structural integrity of the internal lipid membranes.

3.3.3 Transcription components

Transcription-related proteins constitute the largest class of predicted non-structural proteins present in the CroV virion. The packaged transcription machinery consists of eight DNA-directed RNA polymerase II subunits (Rpb1, 2, 3/11, 5, 6, 7, 9, 10), two predicted early transcription factors (VV D6-like TF and VV A7-like TF), transcription elongation factor TFIIS, two VV D11-like TFs, a trifunctional mRNA capping enzyme, polyadenylate polymerase catalytic subunit, DNA topoisomerase IB, and an RNA helicase (Table 3.1). CroV may thus be able to initiate gene transcription immediately upon penetration of the host cell.

The Rpb2 gene contains a 146 aa long intein insertion in its RNA polymerases beta chain signature motif (see section 2.3.8). This intein (like the other three CroV inteins) exhibits all the conserved residues necessary for self-splicing, although there are no experimental data demonstrating that this intein is functional and can remove itself from the host protein (= extein) after translation. Proteomic analysis by LC-MS/MS would be a suitable method to answer this question, if peptides spanning the splice junction were detected. A peptide matching the intein sequence would prove that the intein is packaged, either in its integrated or in its spliced form. A peptide containing part of the extein and part of the intein would be evidence that no protein splicing has occurred. A peptide containing extein sequences from both the N-terminal and C-terminal splice junctions without amino acids from the intein would prove that the extein parts are spliced together correctly and that the intein is functional. Despite 281 peptide spectra and 52% coverage of the Rpb2 protein in this study, no peptides were found that crossed the splice sites or contained any intein residues (Fig. 3.2). Although the intein sequence represents the longest continuous region of the precursor Rpb2 for which no peptide spectrum was found, potentially indicating the absence of the intein, there is no direct evidence for successful intein splicing either, because the theoretical tryptic peptide spanning the splice junction (FSTK) would have been too short to be unambiguously identified.

Figure 3.2 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.

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

3.3.4 DNA repair enzymes

Interestingly, six of the identified virion proteins are predicted DNA repair enzymes. Among them are two putative DNA photolyases, a CPD class I photolyase, and a photolyase-like protein that is phylogenetically related to a recently discovered class of bacterial and euryarchaeal photolyases/cryptochromes (see section 2.3.3). Thus, DNA photodamage could be repaired inside the CroV particle. Such a case of intra-virion DNA photorepair has been demonstrated in Fowlpox virus [234] and could therefore exist in other viruses. With the exception of AP endonuclease, all enzymes of the predicted base excision repair (BER) pathway in CroV are packaged: formamidopyrimidine-DNA glycosylase, family X DNA polymerase, and NAD-dependent DNA ligase. The low RA values of these proteins (0.06-0.26) suggest that only few copies of these enzymes are present in the virus particle. It is therefore possible that the missing BER enzyme, AP endonuclease, was also present, but below the detection limit of this study. The set of virion-located DNA repair enzymes is completed by the CROV200 protein, which contains a 3'-5' exonuclease domain that may be involved in DNA mismatch repair. This is the most extensive DNA repair machinery ever found in a virus particle. Functional studies will be necessary to further characterize these enzymes and elucidate their role in the CroV virion.

3.3.5 Redox proteins

The CroV virion contains three predicted thiol oxidoreductases and two protein disulfide isomerases (PDIs). Enzymes of the PDI family are able to form and break disulfide bonds in other proteins and are involved in a variety of cellular processes, such as oxidative protein folding and antigen presentation [190]. 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 [235]. 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 [236]. 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 genome from the membrane-bound virion core. However, it seems more likely that these redox proteins assist in the folding of structural proteins within the cytoplasmic virion factory. Oxidative protein folding normally occurs in specialized cellular compartments, such as the lumen of the endoplasmic reticulum (ER) and the intermembrane space of mitochondria. The cytosol, on the other hand, is a reducing environment and cysteine residues exposed to the cytosol will preferentially exist in their reduced form, i.e. sulfhydryl groups. A virus that replicates and assembles its virions in the cytoplasm must therefore find a way to catalyze the formation of disulfide bonds in this reducing environment.

Members of the NCLDV clade are especially interesting to study cytosolic disulfide bond formation, because their particles are assembled in large cytoplasmic virion factories, spatially separated from the ER, mitochondria, and the intracellular membrane system. Studies in vaccinia virus revealed a new cytosolic redox pathway for protein disulfide bond formation, which is catalyzed by three viral oxidoreductases, E10, A2.5, and G4 [237]. The oxidizing potential in vaccinia virus is transferred from the ERV1/ALR-family thiol oxidoreductase E10 [238] via A2.5 to the thioredoxin G4, which in turn catalyzes disulfide bond formation in various viral substrates. A comparison of the poxvirus redox components to predicted redox proteins encoded and packaged by CroV and Mimivirus suggests that analogous pathways may exist in these two giant viruses. The VV E10 protein shares conserved residues with the thiol oxidoreductases R596 in Mimivirus and CroV proteins CROV143 and CROV444, all of which are predicted members of the ERV1/ALR family of FAD-binding sulfhydryl oxidases (Fig. 3.3). The most likely homologues of VV thioredoxin G4 are the predicted thioredoxins MIMI_R443 and CROV379 (Fig. 3.4).

Figure 3.3 Sequence alignment of Erv1/Alr-family thiol oxidoreductases.

The catalytic pairs of cysteines are shaded in black; boxed residues have been shown to participate in FAD binding [239]. GenBank accession numbers are as follows: *Saccaromyces cerevisiae* Erv2p, NP_015362; VV E10, YP_232948; Mimivirus R596, YP_003987112.

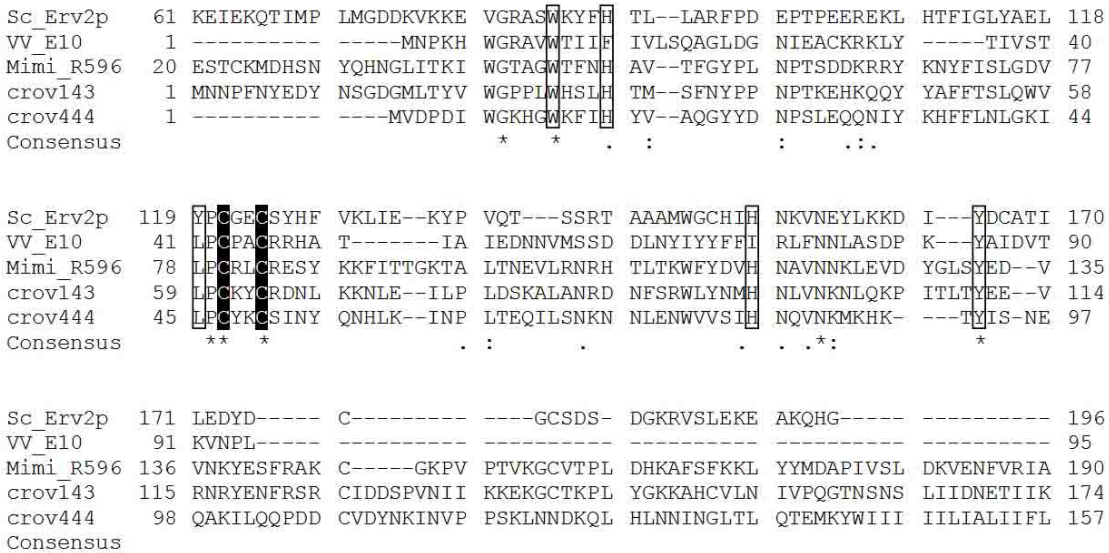
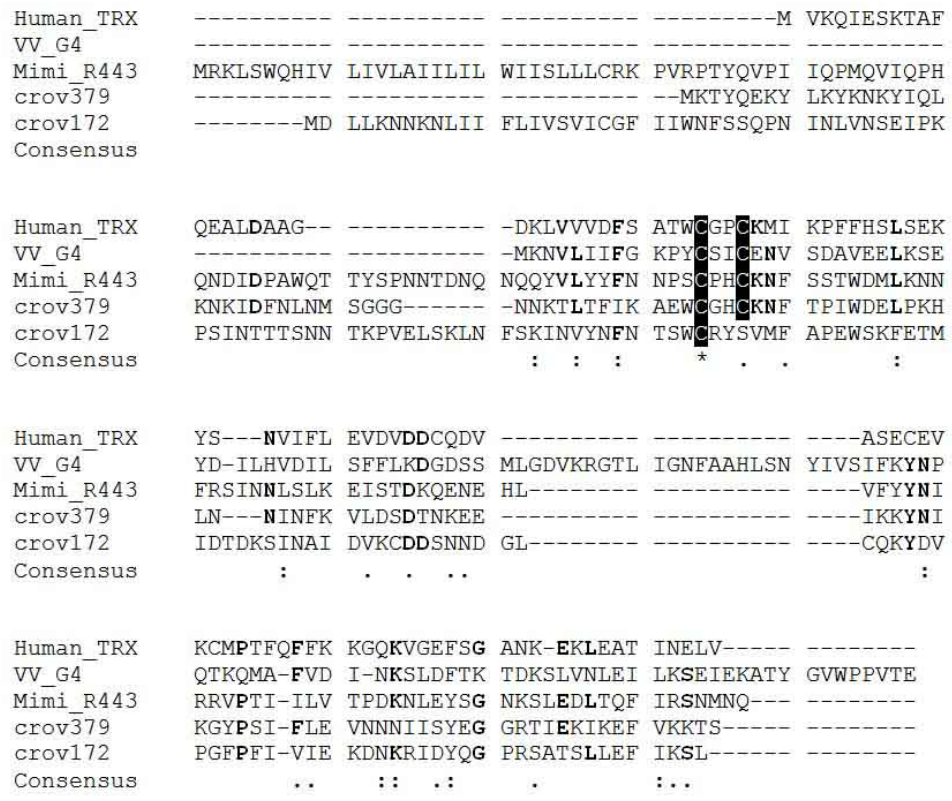


Figure 3.4 Sequence alignment of thioredoxins related to the vaccinia virus G4 protein.

The catalytic pairs of cysteines are shaded in black; residues conserved in at least three organisms are shown in boldface type. GenBank accession numbers are as follows: Human TRX, NP_003320; VV G4, YP_232963; Mimivirus R443, YP_003986950.



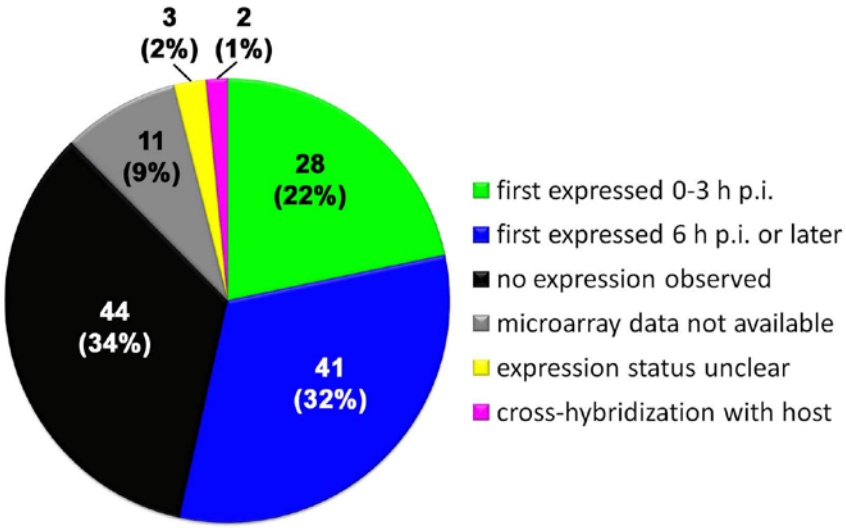
The thioredoxin-like protein CROV172 is also similar to the VV G4 protein, but has only one cysteine in its active site and may thus have reduced or missing redox activity (Fig. 3.4). No potential homologues of the VV A2.5 protein were found in either CroV or Mimivirus; however, it is possible that one of the remaining packaged redox proteins in these viruses may have adopted the role of VV A2.5. The finding that homologous components of a well characterized poxvirus redox pathway are also packaged in CroV and Mimivirus, strongly suggests that these giant viruses experience similar constraints, which requires the catalytic ability for protein disulfide bond formation within the virion. Furthermore, it can be assumed that this conserved viral redox pathway was already present in the common ancestor of CroV, Mimivirus, and the poxviruses, which are members of the nucleocytoplasmic large DNA virus (NCLDV) group. It has recently been proposed that the NCLDV ancestor may already have been endowed with the ability of disulfide bond formation, and that viral redox pathways may have predated oxidative protein folding in eukaryotic organelles [240], since the origin of the NCLDVs is proposed to coincide with or even predate the emergence of eukaryotes [52].

3.3.6 Comparison with microarray data

A temporal gene expression study using a custom-designed DNA microarray revealed that at least 274 CroV genes are transcriptionally active during the infection cycle (see section 2.3.10). The proteome data offers an opportunity to evaluate the microarray results from a different angle. For example, viral proteins detected in the virion must have been transcribed and translated during infection. The proteome data can thus be used to validate the microarray results. Furthermore, virion proteins may be preferentially expressed at a certain period of the infection cycle, e.g. during the early or late stage of the virus life cycle. Figure 3.5 summarizes the gene expression data for CroV virion proteins. Of those virion proteins whose genes were found to be expressed, 28 were first expressed during the early infection stage (0-3 h p.i.) and 41 were first expressed during the late stage (6 h p.i. or later). All structural genes identified in the microarray study were first transcribed during the late stage (Table 3.1). Conversely, 17 of the 41 functionally annotated non-structural virion proteins were expressed early and only 8 were expressed late. This indicates that genes encoding structural virion components are transcribed during the late stage of CroV infection. Among the virion proteins of unknown function, 11 were first expressed during the early stage, and 27 during the late stage. Remarkably, no transcriptional activity was found for 44 confirmed virion components. Clearly, these genes must have been expressed and accordingly, the microarray study must have missed these transcripts. One possible explanation is that the expression levels of these genes were too low to be detected in the microarray study. Although this hypothesis cannot be tested

thoroughly, a rough measure of protein abundance is given by the RA values. Virion proteins that had not been found in the microarray study had an average RA value of 1.13 (ranging from 0.04 to 5.37, lowest and highest 5% excluded), compared to an average RA of 1.76 for virion proteins with confirmed expression (ranging from 0.06 to 11.53). These numbers suggest that low expression levels are unlikely to have caused the observed discrepancy between transcriptomic and proteomic results. Other reasons may lie in preferential degradation of certain mRNA species, quantitative bias introduced during cDNA synthesis, PCR amplification, and fluorescent labelling of cDNAs, or insufficient hybridization of labelled cDNAs with the immobilized oligonucleotide probes. Regardless of the reasons, the LC-MS/MS analysis clearly showed that the microarray analysis underestimated the true extent of transcriptional activity in CroV. Interestingly, 35% of all CroV genes tested in the microarray study were found to be transcriptionally silent (see section 2.3.10), which is almost equivalent to the 34% of confirmed virion proteins with undetectable levels of gene expression. This suggests that nearly all 544 predicted CroV genes could be expressed during infection, which would agree with recent transcriptomic studies in Mimivirus and PBCV-1 [69, 208].

Figure 3.5 Time-resolved transcriptome data of CroV virion proteins.
Numbers indicate how many CroV CDSs with packaged gene products belonged to each category.



3.3.7 Comparison with CroV promoters

Two distinct promoter motifs have been identified in CroV (see section 2.3.11). Genes transcribed early during infection are mostly associated with the "AAAAATTGA" sequence motif, which is located about 40-50 nt upstream of the start codon (Fig. 2.24). Another motif,

consisting of the conserved tetramer "TCTA" flanked by AT-rich regions, is located near nucleotide position -15 and is linked to genes that are first transcribed during the late stage of CroV infection (i.e. at 6 h p.i. or later). Forty-nine (38%) of the 129 genes whose products we found to be packaged are associated with the late promoter motif (Table 3.1); only 12 proteins (9%) are associated with the early promoter motif and all of them are predicted non-structural proteins (seven RNA polymerase subunits, TFIIS, DNA polymerase X, ADP-ribosylglyco-hydrolase, a methyltransferase, and an amino acid transporter). Therefore, all virion proteins of unknown function are either associated with the late promoter motif (36 cases) or are encoded by genes for which no promoter motif could be identified (42 cases). Considering that 8 of the 10 genes encoding structural proteins are associated with the late promoter motif, it can be inferred that many of the virion proteins of unknown function are likely to play a structural role.

3.3.8 CroV and Mimivirus share a conserved set of virion proteins

Virion proteomes of other NCLDV members have been elucidated, including VV [225, 230-232], Singapore grouper iridovirus [222], Mimivirus [68], and the phycodnaviruses EhV-86 [224] and PBCV-1 (D.D. Dunigan and J. Van Etten, personal communication). Of these viruses, only VV, PBCV-1 and Mimivirus virions have a similarly complex protein composition as CroV. At least 7 of the 80 VV proteins that are part of the mature virion [225, 231, 232] have homologues in CroV, and, except for the A18 helicase, are also found in the CroV proteome. In addition to the redox proteins E10 and G4, which are probably functional homologues of CroV proteins CROV143 and CROV379 (despite a lack of significant sequence similarity), the virions of CroV and VV contain at least six homologous proteins. These are the major core protein P4, RNA polymerase subunits 1 and 2 (VV J6 and A24), and transcription factors A7, D6, and D11. Therefore, CroV and VV virions overlap mostly by transcription-related proteins, which constitute 25% of the proteins in the VV mature virion [225].

Proteomics of Mimivirus particles using LC-MS/MS and MALDI-TOF-MS led to the identification of 114 viral proteins [68]. Re-evaluation of this dataset by Claverie *et al.* [59] added 23 statistically well supported proteins to the list, raising the total number of Mimivirus virion proteins to 137. I compared these 137 Mimivirus proteins to the 129 CroV proteins identified in this study. Using a BLASTP E-value cutoff of 10^{-5}, 52 of the 129 CroV CDSs have at least one significantly similar sequence in the Mimivirus genome, and 33 of the 129 CroV virion proteins are significantly similar to a virion protein from Mimivirus (Table 3.1). Although the DNA topoisomerase IB proteins of CroV and Mimivirus do not share significant sequence similarity, they can be considered functional homologues (see section 2.3.4) and were therefore included in the following comparative analysis. The predicted hypothetical protein crov330 is encoded by

two adjacent genes in Mimivirus, which results in a slight difference in the number of CroV-Mimivirus homologues, depending on the virus viewpoint. I here define the number of shared proteins as the number of CroV proteins that have homologues in Mimivirus. Hence, the virion proteomes of CroV and Mimivirus overlap by 34 proteins. Figure 3.6 shows a side-by-side comparison of the virion proteomes of CroV and Mimivirus. Of the 34 shared proteins, only 7 have no functional annotation, which is remarkable considering that 64% of the CroV virion proteins and about half of the virion proteins in Mimivirus have no known function. This implies that most of the packaged proteins of unknown function in CroV and Mimivirus are lineage specific. They may represent structural components that could be required to build the different architectures of the CroV and Mimivirus virions, or they may play a role in virus-host interaction, such as determining host specificity or modulating host defence.

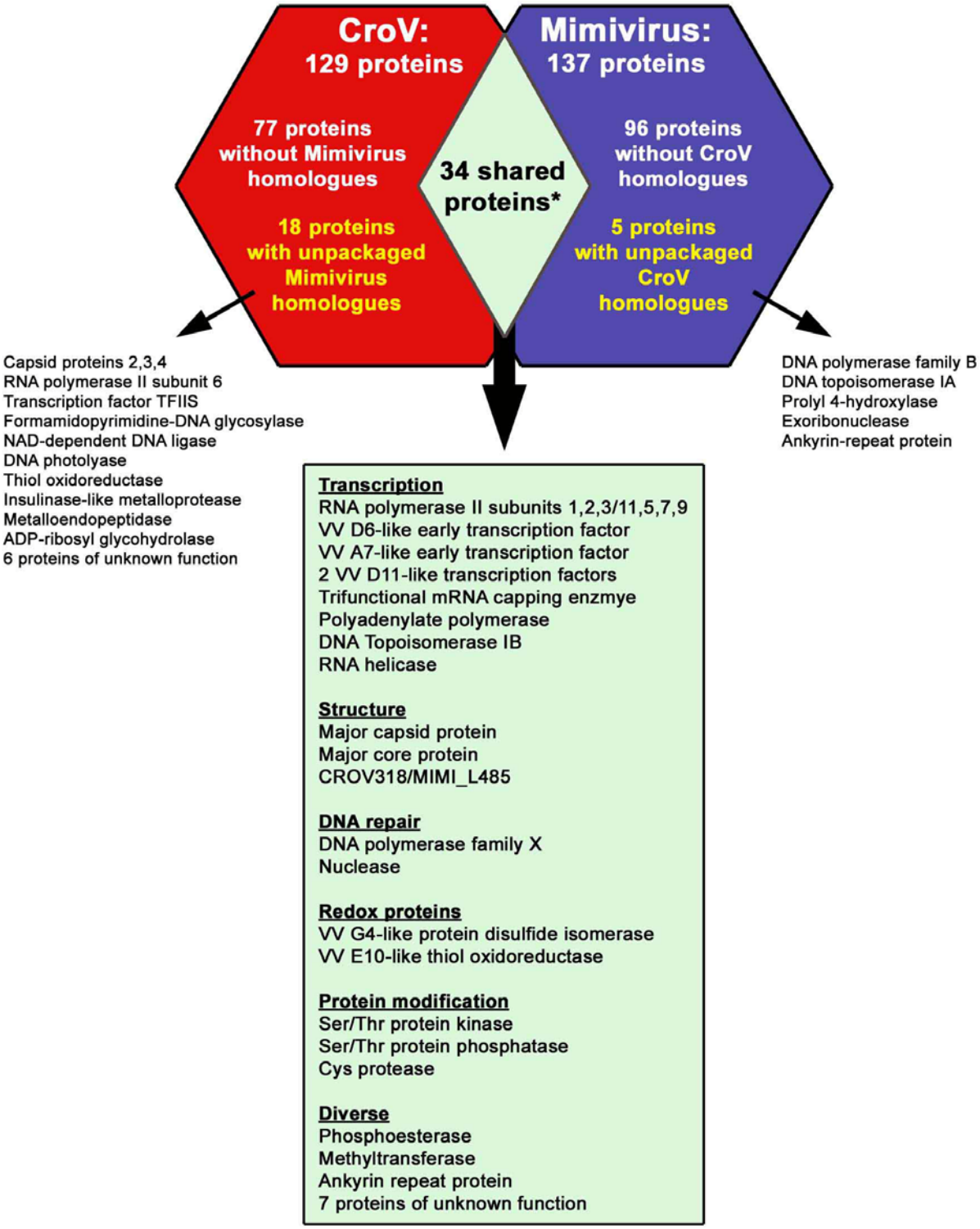
It is significant that the virions of these related, yet so different giant viruses package functionally similar enzymes. Both virions contain a large transcription machinery that consists of a multi-subunit RNA polymerase, early transcription factors, DNA topoisomerase IB, and enzymes for mRNA capping and polyadenylation. They package several DNA repair enzymes including a nuclease and a family X DNA polymerase. The prokaryotic photolyase encoded and packaged by CroV is homologous to two ORFs in Mimivirus that are separated by a predicted transposase, but none of these Mimivirus proteins were found to be part of the virion. Therefore, the Mimivirus photolyase was probably inactivated by insertion of the transposase. Mimivirus and CroV virions share two redox proteins, three enzymes involved in protein modifications (kinase, phosphatase, protease), and further enzymes such as a phosphoesterase and a methyltransferase. The shared proteome is completed by seven proteins of unknown function and three structural proteins, the major capsid protein, the major core protein, and one of the four newly predicted structural proteins in CroV, CROV318, with its Mimivirus homologue L485.

3.3.9 Proteins from Mavirus and phage307

CroV is associated with a small virus parasite, the Mavirus virophage (see chapter 4). The Mavirus virion is ~60 nm in diameter and should not have been retained by the 0.2 µm TFF membrane during CroV purification. However, some Mavirus virions likely ended up in the CroV sample used for LC-MS/MS analysis, as did phage307, which infects an unknown marine bacterium. The full genome sequences for these viruses are available (see chapter 4 and Appendix E) and the peptide mass spectra obtained in this study were searched against a database that also contained ORFs from these two viruses. Ten spectra matched Mavirus CDS18, which encodes the predicted major capsid protein of the virophage (Table 3.1).

Figure 3.6 The shared virion proteome of CroV and Mimivirus.

* Counting CroV proteins. A homologue of crov330 is encoded by two Mimivirus genes, the 34 CroV proteins therefore correspond to 35 Mimivirus proteins.



Phage307 had two statistically supported protein hits, a predicted 37 kDa protein which most likely represents the major capsid protein, and a 14 kDa protein of unknown function. The virions of Mavirus and phage307 are likely to contain further proteins, but since these viruses were not specifically targeted by this proteome study, and were therefore only present as minor contaminants, one would not expect to find minor virion proteins.

3.4 Conclusions

This chapter describes the protein composition of the CroV particle. With 129 packaged proteins, the virion complexity of CroV is rivalled only by Mimivirus. Analysis of the virion proteome in the light of genomic and transcriptomic data of CroV, as well as comparison of this dataset with other virion proteomes, especially with Mimivirus, resulted in valuable new information about the biology and evolution of CroV.

Virion proteomics revealed that protein packaging in CroV and other giant viruses is likely to be a selective and non-random process, since the virions of VV, Mimivirus and CroV contain a select set of enzymes such as transcription and DNA repair proteins. Whereas the targeted incorporation of structural components such as capsid and core proteins might be explained by self-assembly (although this seems unlikely given the complex architecture of giant virus particles and unique capsid features such as the stargate opening in Mimivirus), the mechanism of molecular sorting for non-structural proteins such as RNA polymerases and DNA repair enzymes is completely unknown. Similarly, the intra-virion location of non-structural as well as structural proteins, including the 33 virion proteins with predicted transmembrane domains, remains unknown.

Integration of transcriptomics and proteomics suggests that a previous microarray study on CroV gene expression did not uncover the true transcriptional potential of CroV and that almost all predicted CroV genes may be expressed during infection. CroV genes that encode predicted structural proteins are preferentially expressed from 6 h p.i. onwards and contain the late promoter motif (TCTA) in their 5' upstream region.

The non-structural proteins packaged by CroV fall into distinct categories: transcription, DNA repair, disulfide bond formation, and protein modification. It can therefore be assumed that these enzymatic functions are crucial for the CroV infection strategy. The discovery of several virion-located DNA repair enzymes suggests that DNA damage can be repaired during the extracellular phase of the virus akin to photorepair in poxvirus particles [234], or immediately upon virus entry into the host cell. The extensive array of CroV-encoded transcription

components present in the virion strongly suggests that early transcription in CroV occurs within the confinements of the inner virion core. Such a cytoplasmic intra-core transcription is known only from poxviruses and Mimivirus [54, 241, 242]. For this purpose, the poxvirus VV packages ~22 transcription-related proteins including 8 RNA polymerase subunits [243], and Mimivirus particles contain a similarly comprehensive set of least 14 transcriptional components including 6 RNA polymerase subunits [59, 68]. Likewise, CroV packages 7 RNA polymerase subunits and 9 other transcription proteins and should therefore be capable of carrying out early gene transcription in its core.

Finally, this study revealed that CroV and Mimivirus, the two largest and most complex viruses known, share a significant fraction of their virion proteomes. Although a third of the CroV CDSs have homologues in Mimivirus, the two viruses differ markedly in genome and particle size, genome organization, host specificity (CroV infects a chromalveolate, Mimivirus a unikont), and natural environment (CroV was isolated from the Gulf of Mexico, Mimivirus from the freshwater of a cooling tower). It is therefore highly significant that CroV and Mimivirus contain a common set of 34 virion proteins. Even more remarkably, 27 of the 34 shared proteins have well-characterized homologues in cells or other viruses, which is a rather unusual condition given that typically less than 35% of new viral sequences display similarity to previously described genes, and even less have a putative function [244]. The shared set of virion proteins includes redox proteins, DNA repair enzymes, and a large set of transcription enzymes. This implies that the common ancestor of CroV and Mimivirus, and perhaps even the NCLDV ancestor, already possessed a virion that was endowed with these biochemical functions. A speculative projection of this idea leads to the hypothesis that oxidative protein folding might be a viral invention [240]. The ancient evolutionary origin of the virion proteome is further corroborated by the absence of gene products crov241-crov282 from the CroV virion proteome. These CDSs are located within a 38 kbp genomic region that displays neither early nor late promoter motifs, encodes sugar-modifying enzymes with bacterial affiliation, and lacks homologues in Mimivirus. This suggests that the 38 kbp region was a relatively recent addition to the CroV genome. The integration of these genes, probably from a bacterial source, must have taken place after the CroV lineage had split from the Mimivirus lineage, and probably occurred after the molecular sorting system for virion proteins in giant viruses had evolved.

4 Mavirus, a Novel Virophage Associated with CroV[14]

4.1 Synopsis

DNA transposons are mobile genetic elements that have shaped the genomes of eukaryotes for millions of years by creating new alleles, influencing gene regulation and rearranging chromosome structure [245], yet their origins remain obscure. In this chapter, I report the discovery of a virophage that, based on genetic homology, represents an evolutionary link between double-stranded DNA viruses and *Maverick/Polinton* transposable elements (MP TEs) of eukaryotes. The Mavirus virophage parasitizes the giant *Cafeteria roenbergensis* virus and encodes 20 predicted proteins, including a retroviral integrase and a protein-primed DNA polymerase B. Based on the biology of virophages and the genetic similarity of the Mavirus virophage and MP TEs, the data presented here provide persuasive evidence that MP TEs are derived from an ancient relative of Mavirus, although alternative evolutionary scenarios are possible. These results shed further light on the profound ways in which viruses may have influenced the evolution of cellular life.

4.2 Materials and methods

4.2.1 Culture conditions and virophage purification

Cafeteria roenbergensis strain E4-10 was isolated from coastal waters near Yaquina Bay, OR, USA, as described previously [135]. Mavirus strain Spezl was originally isolated from the same concentrated sea water sample as CroV [101], but remained undetected until 2009, when its genome had amplified sufficiently through serial propagation to produce a contig with 454 pyrosequencing. The existence of Mavirus was visually confirmed by thin-section transmission electron microscopy (see chapter 5). *C. roenbergensis* strain E4-10 culture conditions have been described in section 2.2.1. Cultures were infected with CroV and Mavirus and the resulting lysates were centrifuged for 1 h at 10,500 x g in a Sorvall RC-5C centrifuge (GSA rotor, 4°C) to pellet bacteria and cell debris. The supernatant was subjected to tangential flow filtration (TFF) using a Vivaflow 200, 0.2 µm PES unit (Sartorius Mechatronics, Canada). CroV particles were mostly retained in the concentrate, whereas the Mavirus-containing filtrate was further concentrated in a Vivaflow 200, 100,000 MWCO PES unit. The Mavirus Vivaflow concentrate

[14] A version of this chapter has been published: Fischer M.G. and Suttle C.A. (2011) A virophage at the origin of large DNA transposons. *Science*, March 3, DOI: 10.1126/science.1199412.

was centrifuged for 30 min at 3,000 g in an Eppendorf 5810 tabletop centrifuge to pellet any residual aggregated material. The supernatant was centrifuged for 1 h at 70,000 x g in a Sorvall RC80 ultracentrifuge (SW40 rotor, 20°C). Pellets from two consecutive ultracentrifuge runs were stacked for increased virophage concentration, suspended in a minimal volume of 50 mM Tris-HCl, pH 7.6, and filtered through a 0.1 µm pore size syringe filter (Durapore®, PVDF, 33 mm diameter, Millipore, U.S.A.). The purified and concentrated virophage suspension was used for infection experiments.

4.2.2 Genome sequencing and annotation

The Mavirus genome was sequenced as part of the CroV genome sequencing project, as described previously (see section 2.2.3) [32]. Pyrosequencing on a 454 GS FLX platform (McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada) yielded an 18,973 bp contig with 13-fold coverage. Oligonucleotide primers were designed to test the contig for circular configuration:
forward primer 1 5'-CACAGGATATTTCAGAACCAACGC-3' (T_m=56.3°C);
forward primer 2 5'-TGTATTTAACAATTTAGGAGATTCAG-3' (T_m=49.6°C);
reverse primer 1 5'- TGCTTCTTATGCTTGAAATGTTCC-3' (T_m=54.1°C);
reverse primer 2 5'-GTTTTATTTGGGTTTCCTTCTATAAG-3' (T_m=50.5°C).
Amplifications with different primer combinations were carried out in 25 µl PCR reactions containing 4 ng template DNA, 1 µM of each primer, 1 M betaine, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, and 0.5 units of Platinum Taq DNA polymerase (Invitrogen, Canada). PCR cycle parameters consisted of a single denaturation step for 5 min at 95°C, followed by 35 cycles for 30 sec at 95°C, 1 min at 56°C, and 3 min at 72°C with a temperature rate of 1°C/sec, and a final extension step for 12 min at 72°C. Resulting PCR products were sequenced directly. The complete genome sequence was annotated using Artemis software v12.0 [145]. Functional predictions for CDSs were made on the basis of Pfam and BlastP searches, combined with multiple sequence alignments created with T-Coffee [246] and Muscle [155] to identify conserved protein sites. MrBayes v3.1.2 was used for phylogenetic analysis with rates=gamma, aamodelpr=mixed settings and tree generation following the majority rule consensus method [158]. Phylogenetic trees were displayed using FigTree v1.3.1[15]. Sequences for phylogenetic analysis were chosen in order to represent the taxonomic spectrum of the respective protein. The Mavirus promoter motif was identified by visual inspection of aligned intergenic regions and statistically verified using MEME (E-value $8x10^{-9}$) [153]. BioEdit v7.0.5.3 [247] was used to visualize the alignments in Figs. 4.2 and 4.7. Isoelectric points were predicted using the

[15] http://tree.bio.ed.ac.uk/software/figtree

Greengene software[16]. The genome sequence of Mavirus strain SpezI was deposited in GenBank under the accession number HQ712116.

4.2.3 Co-infection experiments

Triplicate 50 ml cultures of exponentially growing *C. roenbergensis* cells) in modified f/2 enriched sea water medium (see section 2.2.1) were infected with either 50 µl of 50 mM Tris-HCl, pH 7.6 (mock infection), 50 µl of CroV lysate, 50 µl of CroV lysate + 25 µl of concentrated Mavirus suspension, or 25 µl of concentrated Mavirus suspension. Cell densities were monitored until 8 days p.i. by staining 100 µl of culture with 10 µl of Lugol's Acid Iodine solution and counting cells on a hemocytometer. The concentration of CroV particles was measured by epifluorescence microscopy (EFM) after staining with SYBR Gold, as described in detail by Suttle & Fuhrman [144]. Briefly, 10 µl of culture were mixed with 20 µl of 25% EM-grade glutaraldehyde (Canemco, Canada) in 970 µl of dH_2O and incubated for 20 min at 4°C. 800 µl of the fixed sample were then filtered onto a 0.02 µm pore size, 25 mm diameter Anodisc Al_2O_3 filter (Whatman, Canada) and the filter was stained for 17 min in SYBR Gold (Invitrogen, Canada). Stained filters were mounted onto glass microscope slides and examined on an Olympus AX70 microscope under a blue filter. Fluorescent CroV particles were easily distinguishable from bacteria and flagellate cells due to their brightness and symmetrically round shape. Mavirus virions were too small to be counted by EFM under these conditions.

4.2.4 Quantitative PCR

To analyze Mavirus genome replication, quantitative real-time PCR (qPCR) targeting a 159 bp region of the Mavirus DNA polB gene was performed with DNA extracted from infected and mock-infected cultures. Triplicate cultures containing 50 ml of exponentially growing *C. roenbergensis* ($9x10^4$ cells/ml) in f/2 enriched sea water medium were either mock-infected (+25 µl of 50 mM Tris-HCl, pH 7.6), infected with 25 µl of concentrated Mavirus suspension, or infected with 50 µl of CroV lysate + 25 µl of concentrated Mavirus suspension. Subsamples of 200 µl were taken 5 min before infection, 5 min post infection (p.i.), 6 h p.i., and 24 h p.i., and DNA was extracted from these samples following the "Blood or Body Fluids" protocol of the QIAamp DNA Mini Kit (Qiagen, Mississauga, ON, Canada) with a single elution step in 100 µl nuclease-free H_2O (Invitrogen, Canada). DNA was quantified on a Nanodrop spectrophotometer. A 25 µl qPCR reaction contained 1 µl DNA template, 400 nM of each primer (forward primer Spezl-qPCR-1, 5'-ATGGTGCGACTTTTACTTATG-3', T_m=58°C; reverse primer Spezl-qPCR-2, 5'-AAGGGGTTATATTTTCCATTAG-3', T_m=58°C),

[16] http://greengene.uml.edu/programs/FindMW.html

and 12.5 µl of iQ SYBR Green Supermix (Bio-Rad, Canada). qPCR amplification was carried out on a Bio-Rad Opticon 2 instrument. The amplification protocol consisted of a single denaturation step for 10 min at 95°C, followed by 45 cycles for 20 sec at 95°C, 20 sec at 55°C, and 15 sec at 72°C with a temperature rate of 2°C/sec. PCR amplification with primers Spezl-qPCR-1 and Spezl-qPCR-2 yielded a single product of ~160 bp, as verified by agarose gel electrophoresis. Fluorescence was measured after the extension step of each cycle and a final melting curve was recorded (55°C - 95°C in 0.5°C increments). The C_T values for all no-template controls were above 45 (i.e. out of range).

4.3 Results and discussion

4.3.1 Mavirus replication in *Cafeteria roenbergensis*

During the characterization of *Cafeteria roenbergensis* virus (CroV), a significantly smaller virus associated with infection by CroV was observed and sequenced by 454 pyrosequencing. This virus, named Mavirus (for *Maverick* virus), caused in co-infected cultures a significant decline in CroV production and host cell lysis. As shown in Table 4.1, cultures infected with CroV lysate declined after ~3 days post infection (d p.i.) and yielded $6x10^7$ CroV virions per ml. When a concentrated Mavirus suspension was added together with the CroV inoculum, the flagellate population remained healthy and no CroV particles were detected by epifluorescent microscopy at 8 d p.i.. Addition of the concentrated Mavirus suspension to *C. roenbergensis* cultures in the absence of CroV had no observable effect on flagellate growth (Table 4.1). This indicates that Mavirus is not able to induce lysis in *C. roenbergensis*. The ability of Mavirus to replicate in *C. roenbergensis* in the presence and absence of CroV was examined by quantitative PCR. Table 4.2 shows the threshold cycle (C_T) values for a 159 bp product of the Mavirus DNA polymerase B gene, determined for DNA samples extracted at different time points post infection. C_T values are inversely proportional to the logarithm of the starting quantity of template DNA, i.e. higher initial copy numbers of the template sequence will result in lower C_T values. In *C. roenbergensis* cells co-infected with CroV, Mavirus DNA was replicated, as indicated by the lower C_T values at 24 h p.i. in Table 4.2. In contrast, Mavirus did not replicate in the absence of CroV, as demonstrated by the consistently high C_T values throughout infection. Mavirus was therefore classified as a virophage *sensu* La Scola *et al*. [31], who coined the term "virophage" to emphasize the unique status of Sputnik, a small virus associated with Mimi- and Mamavirus, as a true virus parasite of another virus, which sets it apart from other satellite viruses (see section 1.3.1) [31, 105]. According to this definition, virophages are predicted to have no nuclear phase in their infection cycle, to replicate in the virion factory of their host virus, and to be dependent on enzymes provided by their host virus, rather than by the host cell. Support for this hypothesis comes from electron microscopy data showing Sputnik particles within the Mamavirus factory [31], as well as from the finding that Sputnik shares highly specific regulatory DNA sequences (promoter signals and transcriptional terminators) with Mimivirus [69, 105]. It can therefore be assumed that at least virophage transcription and capsid assembly occur within the virion factory of the host virus. However, further studies on gene expression and DNA replication of these viruses are necessary to validate and refine the virophage concept.

Table 4.1 Cell and virus concentrations during infection of *C. roenbergensis* with CroV and Mavirus.

Cells were counted on a hemocytometer with a detection limit of $1.4x10^3$ cells/ml. CroV particles were counted by epifluorescence microscopy, which had a detection limit of $2.4x10^5$ viruses/ml. Cultures infected with CroV lysate declined at 3 days post infection (p.i.), whereas cultures, to which purified Mavirus had been added, survived. CroV concentrations in Mavirus-infected cultures were below detection limit (BDL).

	-5 min p.i.	**1 d p.i.**	**3 d p.i.**	**8 d p.i.**	
	Cell concentration ($x10^5$ cells/ml)	Cell concentration ($x10^5$ cells/ml)	Cell concentration ($x10^5$ cells/ml)	Cell concentration ($x10^5$ cells/ml)	CroV concentration ($x10^7$ viruses/ml)
Uninfected	0.8 ± 0.1	7.6 ± 1.6	10.5 ± 0.4	37.5 ± 18.2	BDL
+ CroV lysate		2.6 ± 0.4	0.2 ± 0.1	BDL	6.1 ± 1.3
+ CroV + Mavirus		2.3 ± 0.3	9.4 ± 2.1	51.5 ± 32.4	BDL
+ Mavirus only		4.8 ± 0.9	16.8 ± 5.0	18.0 ± 6.4	BDL

Table 4.2 Quantitative PCR analysis of Mavirus replication.

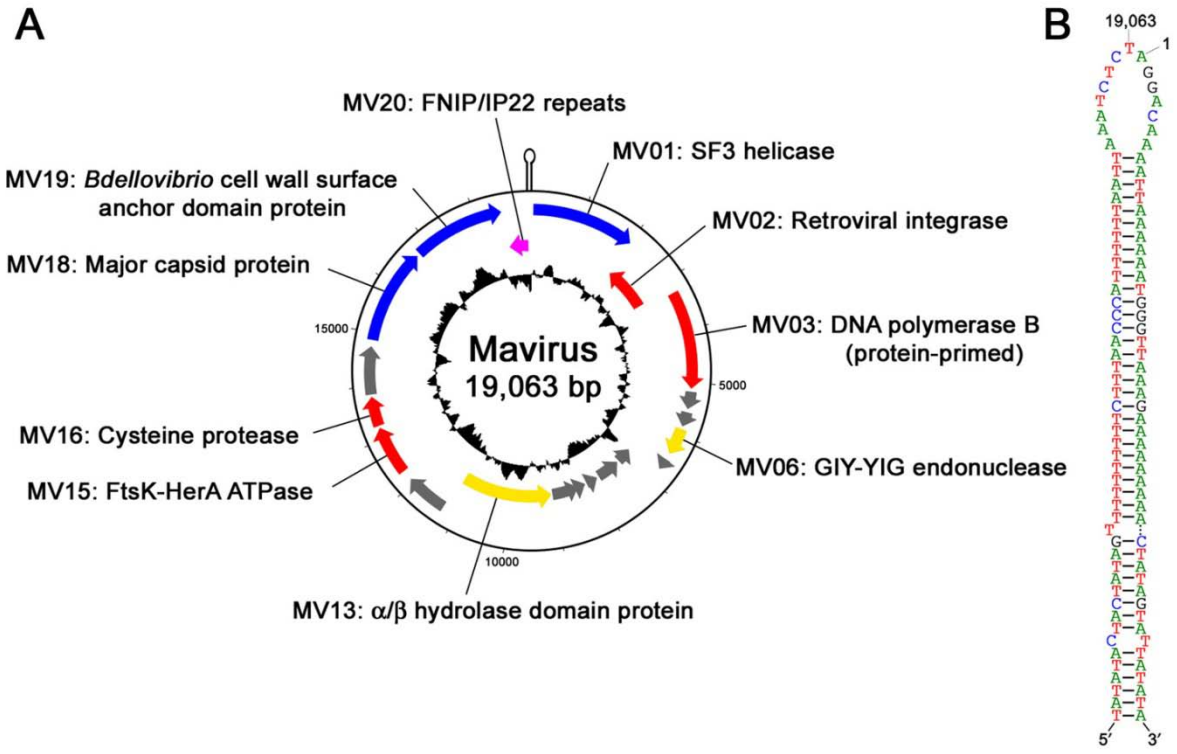
	C_T values -5 min p.i.	**C_T values 5 min p.i.**	**C_T values 6 h p.i.**	**C_T values 24 h p.i.**
Uninfected	32.98 ± 0.77	n/a	34.53 ± 4.20	30.69 ± 0.92
+ Mavirus only		27.81 ± 0.40	27.76 ± 0.32	28.17 ± 0.56
+ CroV + Mavirus		27.11 ± 0.34	27.94 ± 0.52	23.76 ± 2.03

4.3.2 The Mavirus genome

The Mavirus genome is a 19,063 bp circular dsDNA molecule with a GC content of 30.26% (Fig. 4.1A). The genome was found to contain 20 predicted protein-coding sequences (CDSs) with an average length of 883 nt, which resulted in a genome coding density of 92.6%. A nearly perfect inverted repeat of ~50 bp was situated in the intergenic region between CDSs MV20 and MV01. This region has the potential to adopt a hairpin structure with a 43 bp stem and a 15 nt loop (Fig. 4.1B). The adenine at the top of this hairpin loop was designated position 1 of the genome. Analysis of the intergenic regions revealed a conserved promoter motif (Fig. 4.2) preceding all 20 CDSs and that is highly similar to a putative late promoter motif in CroV. This implies that Mavirus gene expression is governed by the transcription machinery of CroV during the late stage of infection.

Figure 4.1 The Mavirus genome.

(**A**) Genome diagram of the Mavirus virophage. Circles from outermost to innermost represent CDSs on the forward strand, the location of promoter motifs, CDSs on the reverse strand, and GC content (relative to the average of 30.26%). Conserved MP TE genes are shown in red; genes that are occasionally encoded by MP TEs are shown in blue. The FNIP/IP22 repeat-containing gene is shown in pink, additional genes with functional annotations are shaded yellow. A hairpin structure indicates the position of the inverted repeat upstream of MV01. (**B**) DNA sequence and predicted secondary structure of the inverted repeat upstream of MV01. The first and last bases of the genome are numbered.



Ten Mavirus CDSs showed sequence similarity to proteins from eukaryotes, bacteria, retroviruses, and dsDNA viruses including at least four genes found in Sputnik (Table 4.3). The physical properties of the predicted Mavirus genes are listed in Table 4.4. MV01 encodes a superfamily 3 helicase (S3H) similar to the D5-like ATPases, which is thought to be the main replicative helicase in NCLDVs [50]. In contrast to NCLDVs, where the S3H domain is located at the C-terminus of a primase-helicase fusion protein [50], the Mavirus S3H domain is located towards the N-terminus and is followed by a domain of unknown function. Remarkably, MV02 encodes an integrase that belongs to the rve superfamily of retroviral integrases (Pfam entry PF00665, E-value $1x10^{-16}$). Retroviral integrases (rve-INTs) are usually expressed as part of the POL polyprotein in retroviruses, but rve-INTs have also been identified in a variety of eukaryotic genomes, where they occur independently of retroelements and have been termed cellular (c-) integrases [248].

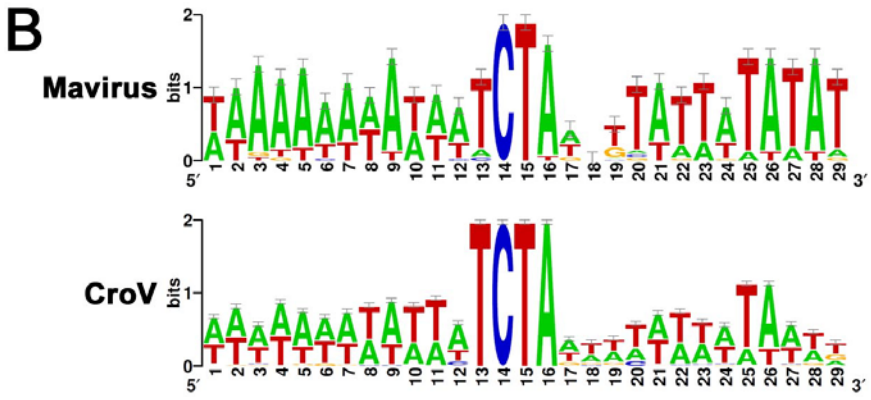
Figure 4.2 The Mavirus promoter motif.

(A) Alignment of the 5' upstream regions of all 20 CDSs. The predicted start codons are printed in uppercase. The bottom line shows the 80% consensus sequence. (B) Sequence logos of the Mavirus promoter motif, created from a gap-free alignment of all 20 promoter regions, and the predicted late promoter motif found in CroV (S3). (C) Comparison of the positional distribution of the 'TCTA' promoter motifs in CroV and Mavirus.

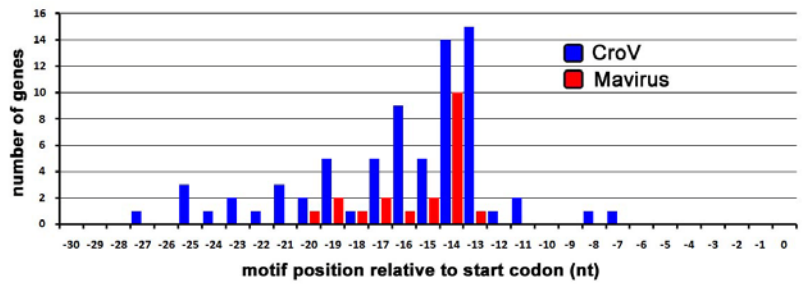
A

```
MV01     aaagaaaaaaaaactatagtat------tatataATG
MV02     tatatattaaattctattata-------tattatATG
MV03     aaaaattatttatctatagta-------tatataATG
MV04     taaataaaaaaatctatattat------tatat-ATG
MV05     ttaaaaatatattctagtttaat-----tatat-ATG
MV06     aaataaataatctctagtctat------tatat-ATG
MV07     ttaaataaataatctttagtat------agtat-ATG
MV08     taaaaaatattatctaagtgtata----tatataATG
MV09     ttaaattaataatctaagttatatta--tatataATG
MV10     taataataattttctaatta--------tttttaATG
MV11     ttaaataaatattctagattat------tatat-ATG
MV12     aaaaaaatattatctaaggcattata--tatataATG
MV13     aaaaatttaatatctaacttatta----tatat-ATG
MV14     ataaacaaaaaacctaaatta-------tattatATG
MV15     ttaaaaataaaatctatctttttata--tataa-ATG
MV16     taaaaaaattattctattgttt------tatat-ATG
MV17     aaaataatatattctaagttta------tatat-ATG
MV18     aaaaaaatattatctaagttaa------tatat-ATG
MV19     taaaaaataatatctaagttatta----tatataATG
MV20     aagaaaaaaaaaactatagtagtatatatataa-ATG
808 con  ..AAA...A...TCTA...T........TATAT.ATG
```

B

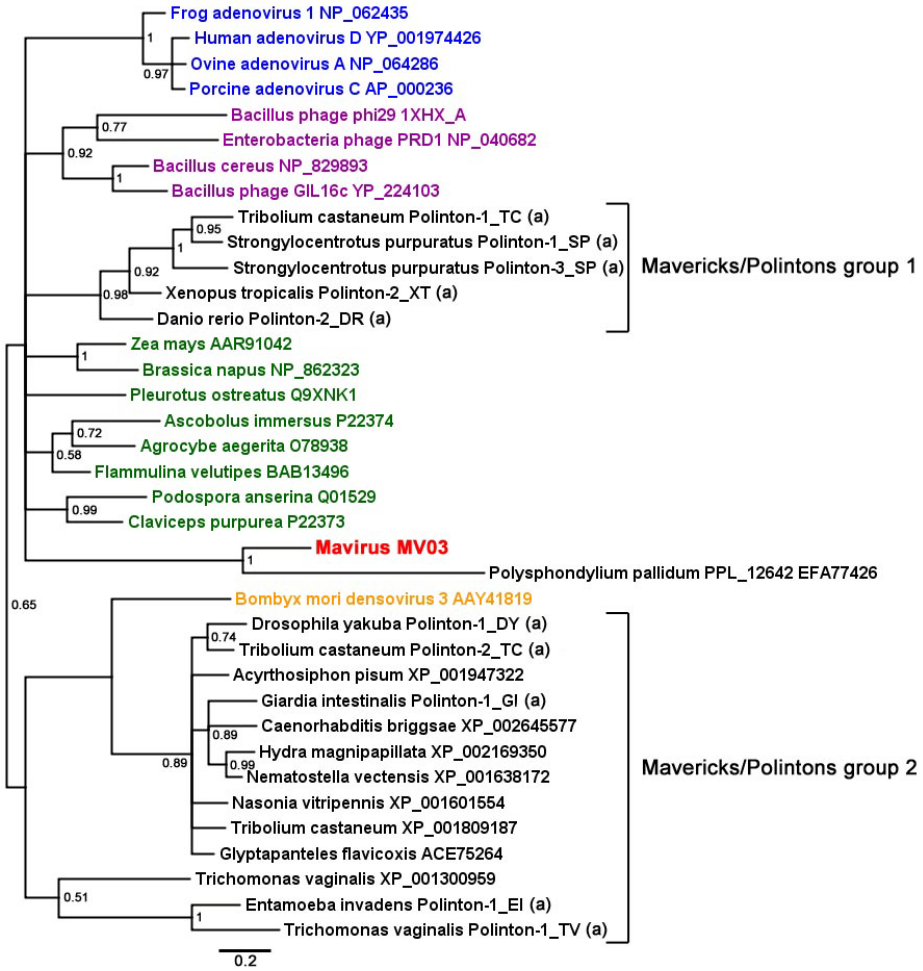


C



Both the Mavirus integrase and the eukaryotic c-integrases contain a C-terminal CHROMO domain (PF00385, E-value $4x10^{-4}$), which is a conserved region of ~60 aa capable of interacting with specific parts of the chromatin [248, 249]. The CHROMO domain might thus be responsible for determining the target specificity of the Mavirus integrase. MV03 encodes a predicted protein-primed DNA polymerase B (DNA_pol_B_2, PF03175, E-value $9x10^{-5}$), with homologues in bacteriophages, adenoviruses, densoviruses, as well as mitochondrial genomes of plants and fungi. However, the closest homologue of Mavirus MV03 was found in the slime mould *Polysphondylium pallidum* (Fig. 4.3).

Figure 4.3 Phylogenetic tree of protein-primed DNA-dependent DNA polymerase B.
The unrooted Bayesian Inference tree was created from an alignment of conserved regions of protein-primed DNA pol B enzymes from selected adenoviruses, bacteriophages, linear mitochondrial plasmids, and MP TEs. Nodes are labeled with posterior probabilities and GenBank accession numbers are listed where available. Adenoviruses are shown in blue, bacteriophages in purple, linear mitochondrial plasmids in green, a sequence from *Bombyx mori* densovirus-3 in orange, and MP TEs in black. Sequences labeled (a) were downloaded from RepBase [250] and MP TEs groups 1+2 correspond to POLB1+2 clades from Ref. [255].



Interestingly, both Mavirus and *P. pallidum* DNA polymerases contain an insertion sequence in their catalytic site that may represent the remnants of an intein, inserted between the amino acid motifs YxG and TDG. With the putative intein sequence spliced out, the YxGTDG sequence corresponds to the YGDTDS catalytic motif found in B-family polymerases [250]. Inteins inserted at the exact same site of this motif have been found in other viruses, such as Mimivirus, CroV, HaV, and CeV (see section 2.3.8) [196, 197]. In contrast to these giant viruses, however, the Mavirus and *P. pallidum* inteins are extremely short with 58 aa and 55 aa, respectively (Fig. 4.4). The absence of the highly conserved Cys or Ser at the N-terminal splice junction and the missing Asn at the C-terminal splice junction suggests that these insertion sequences are incapable of self-excision, at least according to the traditional model of intein splicing [192]. MV06 contains the signature motif of GIY-YIG endonucleases (PF01541, $2x10^{-5}$) and MV13 exhibits an alpha/beta hydrolase domain similar to the ones found in lipases. MV15 bears Walker A, Walker B, and P9/A32-specific motifs, which define the DNA-pumping ATPases of the FtsK-HerA clade that are found in membrane-containing viruses (Fig. 4.5) [251]. The adjacent gene, MV16, encodes a predicted adenoviral cysteine protease that had its closest homologue in the Sputnik virophage (Fig. 4.6). Also similar to Sputnik was the predicted MCP encoded by MV18. The capsid proteins of the two virophages shared no primary sequence similarity to other viruses, and both have been verified by proteomic analysis (see section 3.3.9) [31]. The C-terminal domain of MV19 shows significant similarity to a cell wall surface anchor family protein from *Bdellovibrio bacteriovorus* (BLASTP E-value $3x10^{-21}$, 37% identity over 168 amino acids). Finally, MV20 contains three FNIP/IP22 family repeats (PF05725, E-value $3x10^{-8}$). These ~22 amino acid long repeats are present in Mimivirus, the slime moulds *Dictyostelium discoideum* and *P. pallidum*, the brown alga *Ectocarpus siliculosus*, and in the CroV genome, where they have expanded to more than 400 copies (see section 2.3.9) [204]. It is interesting to note that the FNIP/IP22 repeats are located directly adjacent to the predicted hairpin structure, which might be involved in the initiation of genome replication. Future studies may be able to show if and how these repeats interact with the FNIP/IP22 repeats in the CroV genome.

Table 4.3 Mavirus gene annotations.

CDS	Predicted function or domain	Top BlastP hit (aa identity, alignment length in aa, E-value)	Pfam and NCBI Conserved Domain Database matches (domain ID, description, E-value)
MV01	Superfamily III helicase	Putative D5 N-terminal domain family protein, *Octadecabacter antarcticus* 307 (28%,171, $2x10^{-8}$)	cl11759, phage/plasmid primase, P4 family, C-terminal domain, $7x10^{-8}$
MV02	Retroviral integrase	Hypothetical protein PPL_12641, *Polysphondylium pallidum* PN500 (33%, 341, $1x10^{-31}$)	PF00665, integrase core domain, $1x10^{-16}$; PF00385, chromo domain, $5x10^{-4}$; cl01316, rve superfamily, $9x10^{-15}$
MV03	Protein-primed DNA polymerase B	Hypothetical protein PPL_12642, *Polysphondylium pallidum* PN500 (25%, 343, $6x10^{-17}$)	PF03175 (via NCBI CDD), DNA_pol_B_2 DNA polymerase type B, organellar and viral, $3x10^{-4}$
MV04	-	-	-
MV05	-	-	-
MV06	GIY-YIG endonuclease	Conserved hypothetical protein, *Phytophthora infestans* T30-4 (33%, 161, $3x10^{-13}$)	PF01541, GIY-YIG catalytic domain, $2x10^{-5}$
MV07	-	-	-
MV08	-	-	-
MV09	-	Hypothetical protein BRAFLDRAFT_119045, *Branchiostoma floridae* (29%, 134, 0.13)	-
MV10	-	-	-
MV11	-	-	-
MV12	-	-	-
MV13	α/β hydrolase	Predicted lipase, *Physcomitrella patens* subsp. patens (37%, 85, $3x10^{-6}$)	UPF0227, uncharacterized protein family (containing alpha/beta hydrolase fold), $1x10^{-5}$; cd00519, Lipase_3, $1x10^{-10}$
MV14	-	Putative lipoprotein, *Vibrio parahaemolyticus* 16 (23%, 142, 0.007)	-
MV15	FtsK-HerA DNA packaging ATPase	V3 protein, Sputnik virophage (25%, 227, $1x10^{-4}$)	-
MV16	Cysteine protease	V9 protein, Sputnik virophage (32%, 116, $6x10^{-10}$)	-
MV17	-	Hypothetical protein EUBSIR_00774, *Eubacterium siraeum* DSM 15702 (26%, 118, 0.16)	-

CDS	Predicted function or domain	Top BlastP hit (aa identity, alignment length in aa, E-value)	Pfam and NCBI Conserved Domain Database matches (domain ID, description, E-value)
MV18	Major capsid protein	V20 protein, Sputnik virophage (25%, 183, 0.02)	-
MV19	Similar to cell wall surface anchor domain from *Bdellovibrio bacteriovorus*	Cell wall surface anchor family protein, *Bdellovibrio bacteriovorus* HD100 (37%, 168, $3x10^{-21}$)	-
MV20	FNIP repeats	FNIP/IP22 repeat protein, *Ectocarpus siliculosus* (37%, 120, $8x10^{-15}$)	PF05725, FNIP repeat, $3x10^{-8}$; cl05341, FNIP super family, $5x10^{-4}$

Figure 4.4 Alignment of DNA polymerase B inteins from Mimivirus and CroV, and potential mini-inteins from *Polysphondylium pallidum* and Mavirus.

Identical residues are shaded in red, similar residues are shaded green. Extein sequences are printed in lower case, intein sequences in upper case. Amino acids essential for intein self-splicing are indicated by arrows. GenBank accession numbers are YP_142676 (*Acanthamoeba polyphaga* mimivirus), ADO67531 (CroV), and EFA77426 (*P. pallidum*).

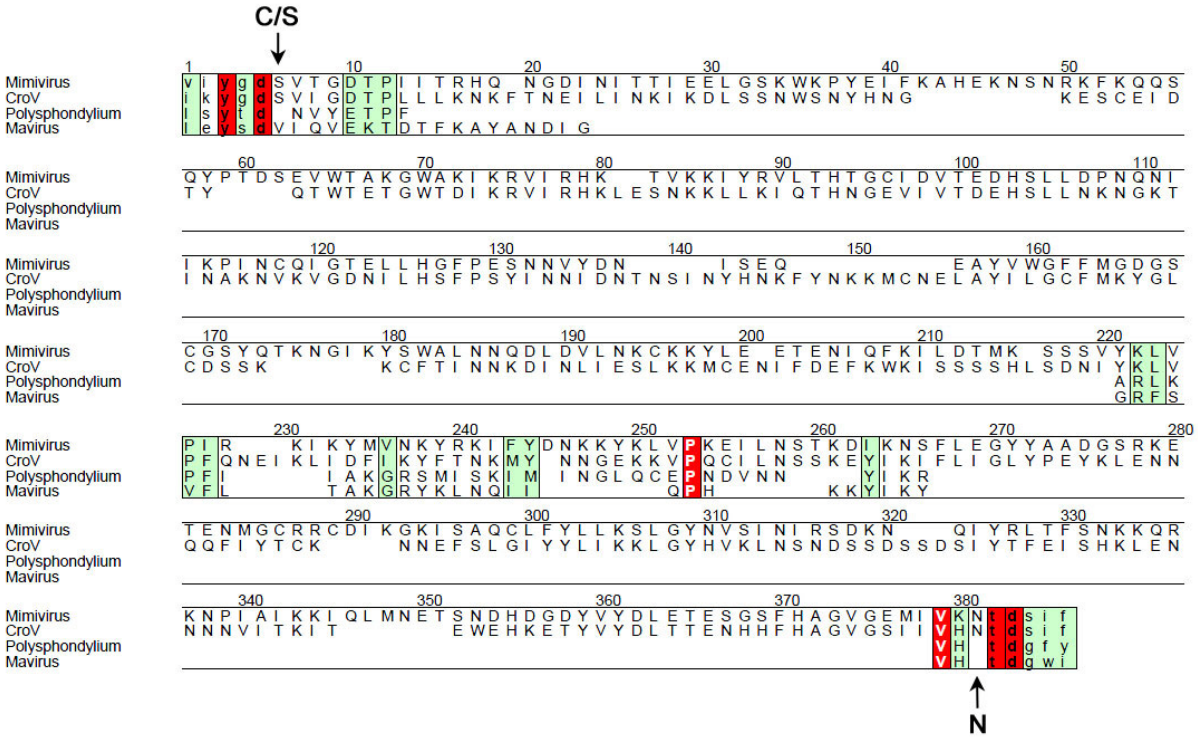


Figure 4.5 Phylogenetic tree of FtsK-HerA-like DNA packaging ATPases.

This unrooted Bayesian Inference tree was created from a 63 aa alignment of conserved regions of DNA packaging ATPases from NCLDVs, bacteriophages, archaeal viruses, and MP TEs. Nodes are labelled with posterior probabilities and GenBank accession numbers are listed where available. NCLDV members are shown in blue, virophages in red, bacteriophages in purple, archaeal viruses in orange, and MP TE sequences in black. Sequences labelled (a) were downloaded from RepBase [252]; sequences labelled (b) were acquired from Pritham *et al.* [253].

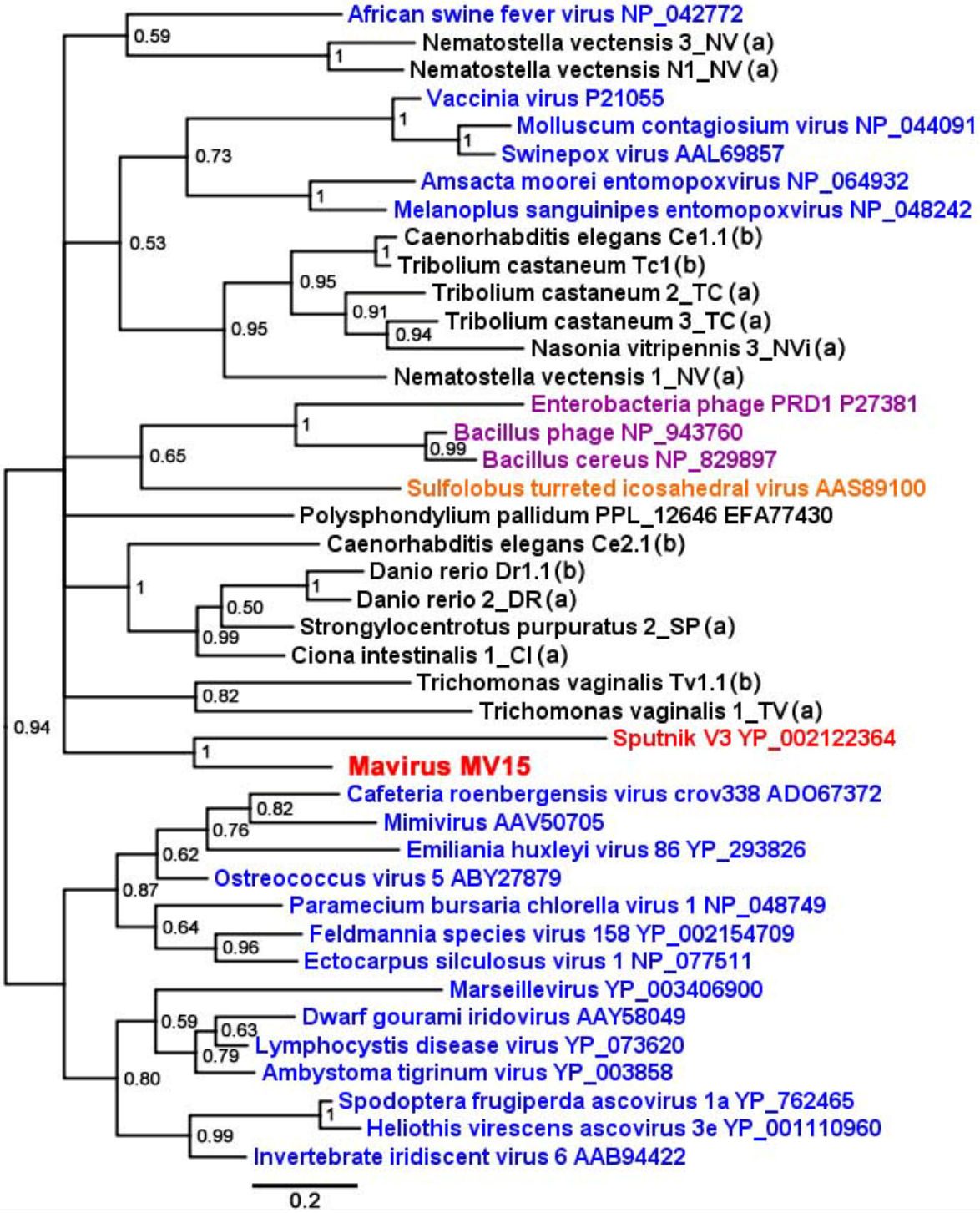


Figure 4.6 Phylogenetic tree of cysteine proteases.

Shown is an unrooted Bayesian Inference tree that was created from a 67 aa alignment of conserved regions of predicted cysteine proteases from selected adenoviruses, NCLDVs, yeast, and MP TEs. NCLDV members are shown in blue, virophages in red, adenoviruses in purple, the yeast ULP1 protein in gray, and MP transposon sequences in black. Nodes are labelled with posterior probabilities and GenBank accession numbers are listed where available. Sequences labelled (a) were downloaded from RepBase [252]; sequences labelled (b) were acquired from Pritham *et al*. [253]. The *P. pallidum* sequence for the predicted cysteine protease PPL_12644 (c) was extracted from the nucleotide sequence of the corresponding contig (GenBank ADBJ01000043).

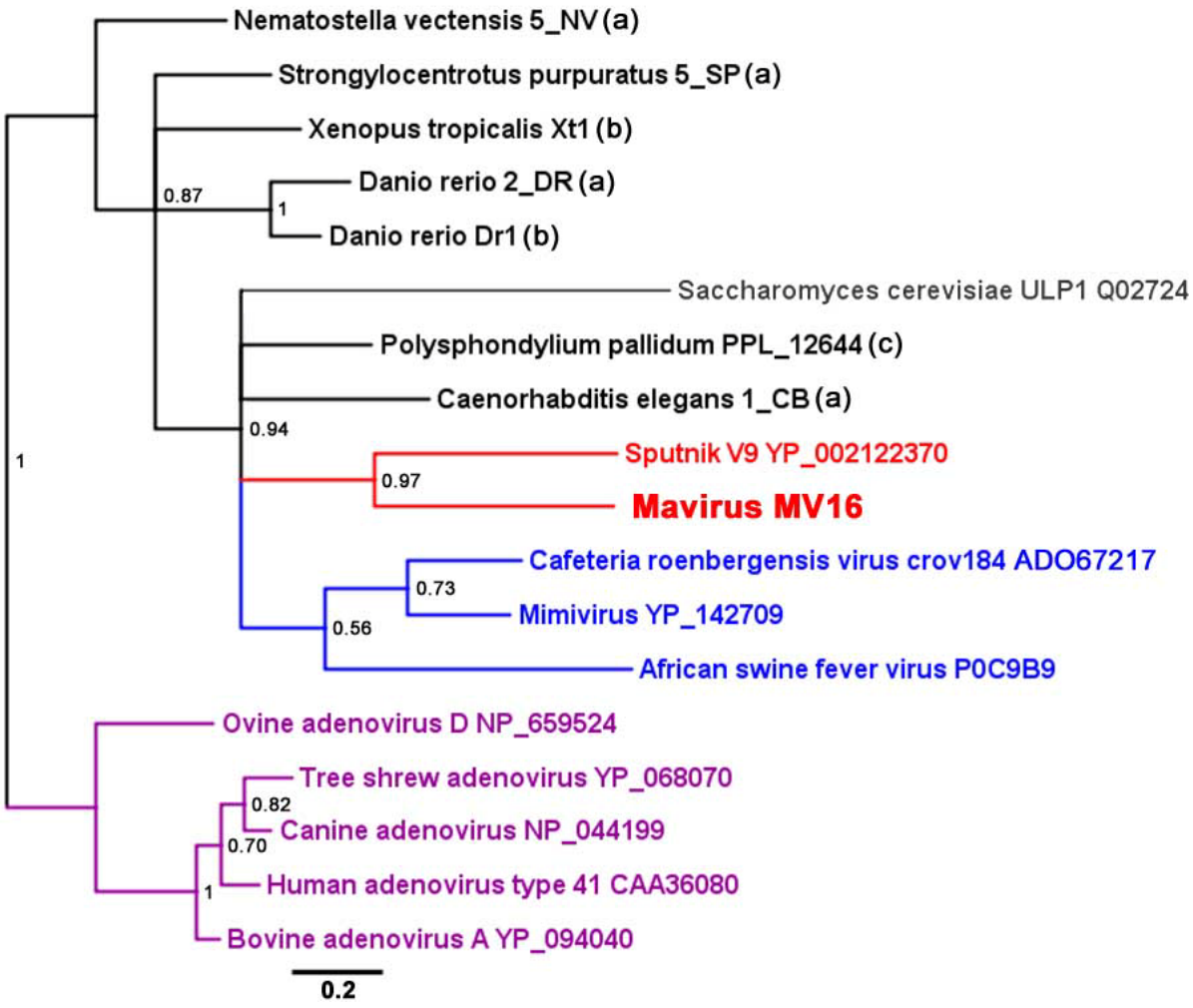


Table 4.4 Physical properties of Mavirus CDSs.

No transmembrane helices or signal peptides/anchors were predicted.

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI
MV01	52-2010	1	31.44	652	75.32	8.10
MV02	3138-2062	5	29.81	358	41.91	8.98
MV03	3249-5102	3	30.31	617	72.47	8.73
MV04	5154-5492	3	31.27	112	13.59	9.48
MV05	559-5833	2	34.74	94	11.02	9.76
MV06	5894-6391	2	30.52	165	19.76	9.23
MV07	6522-6740	3	29.22	72	8.59	10.13
MV08	7180-6812	6	40.65	122	13.40	10.21
MV09	7800-7228	5	28.45	190	21.82	7.65
MV10	8133-7846	5	23.61	95	11.36	6.57
MV11	8337-8173	5	23.64	54	6.94	9.07
MV12	9000-8365	5	33.33	211	24.28	9.33
MV13	11187-9049	5	34.27	712	80.61	9.37
MV14	11262-12077	3	23.65	271	31.92	9.64
MV15	12273-13205	3	28.62	310	36.42	8.87
MV16	13239-13808	3	30.00	189	21.75	6.59
MV17	13850-14761	2	31.03	303	35.23	4.58
MV18	14876-16696	2	35.80	606	66.34	8.84
MV19	16755-18503	3	36.76	582	62.70	4.69
MV20	19005-18547	5	24.62	152	18.24	9.49

4.3.3 Comparison of Mavirus and Sputnik virophages

Sputnik is the first and only virophage that has been described thus far and its discovery raised the question whether similar satellite agents are abundant in nature and what their host range and genetic composition might be. The discovery of Mavirus offers the first opportunity to compare Sputnik to another virophage.

Sputnik has an icosahedral capsid with a diameter of 50 nm as determined by transmission electron microscopy (TEM) [31, 254], or 75 nm as determined by cryo-electron microscopy [104]. The Mavirus virion is also icosahedral and scales up to ~60 nm in diameter in TEM thin sections (see Fig. 5.12). The two virophages therefore have similar capsid morphologies and dimensions. The slightly larger capsid of Mavirus may be due to a minor genome size difference between Mavirus (19,063 bp) and Sputnik (18,343 bp). Both virus genomes are circular and have high AT contents, 69.4% A+T for Mavirus and 73.0% A+T for Sputnik. However, the Sputnik genome lacks the terminal inverted repeats found in Mavirus. The two genomes encode about the same number of genes (20 CDSs in Mavirus and 21 CDSs in Sputnik), but have different coding densities (80.0% Sputnik, 92.6% for Mavirus). The finding that the shorter Sputnik genome with larger intergenic regions encodes more genes than the longer and gene-denser Mavirus genome can be explained by a difference in the average gene length (692 nt for Sputnik and 883 nt in Mavirus). The coding content of the two virophages differs significantly, as shown in Fig. 4.7. Only five genes are homologous between Mavirus and Sputnik, and even those genes exhibit low sequence conservation (Table 4.5). The Mavirus capsid protein is related to the Sputnik capsid protein (Fig. 4.8), which is predicted to adopt the same double jelly-roll fold found in other viruses of the PRD1-adenovirus lineage [104, 123]. The jelly-roll capsid gene is a so-called viral hallmark gene (VHG), because it is found in a wide range of different viruses [48].

Figure 4.7 Gene homologies between Mavirus and Sputnik

In this linear view of the Sputnik and Mavirus genomes, homologous genes are shown in the same colour and are labelled with their predicted function. Abbreviations: ATPase, FtsK-HerA DNA packaging ATPase; MCP, major capsid protein; PR, DNA primase; PRO, cysteine protease; S3H, superfamily 3 helicase; ZF/EN, zinc finger domain/endonuclease.

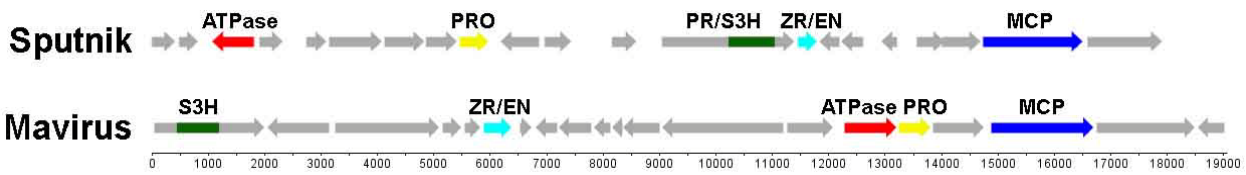


Figure 4.8 Multiple sequence alignments of virophage and MP transposon capsid proteins.

(**A**) Optimized MUSCLE [155] alignment of the Mavirus MV18 protein and the Sputnik V20 capsid protein. The alignment was visualized with BOXSHADE . (**B**) PRALINE [255] alignment of Mavirus MV18, Sputnik V20, and PY proteins found in MP TEs. The shaded regions represent secondary structure predictions using DSSP [256] and PSIPRED [257]; red corresponds to predicted alpha helices, blue to predicted beta strands. GenBank accession numbers are given where available; sequences labelled (a) were downloaded from RepBase [250].

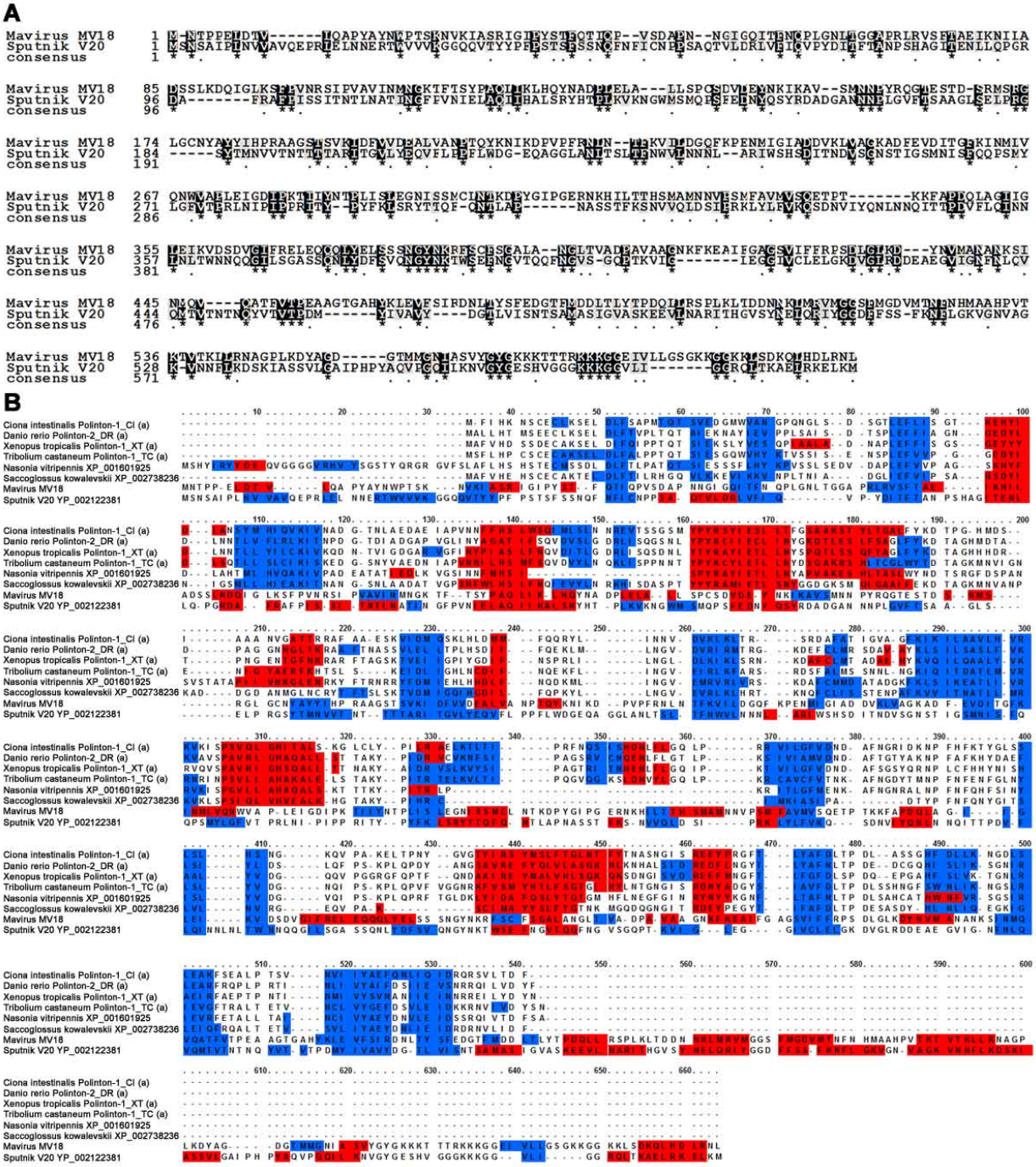


Table 4.5 Sequence similarities of Mavirus and Sputnik proteins as determined by BLASTP analysis.

Mavirus CDS, predicted function	Sputnik CDS, predicted function	BLASTP via bl2seq (E-value, aa identity, alignment length in aa)
MV01, SF3 helicase	V13, primase-helicase	-
MV06, GIY-YIG endonuclease	V14, Zn ribbon- containing protein	$3x10^{-6}$, 30%, 64
MV15, FtsK-HerA packaging ATPase	V3, FtsK-HerA packaging ATPase	$2x10^{-11}$, 26%, 227
MV16, cysteine protease	V9, unknown	$1x10^{-16}$, 33%, 116
MV18, major capsid protein	V20, major capsid protein	$5x10^{-9}$, 26%, 183

Two additional VHGs are found in virophages, an FtsK-HerA-like genome packaging ATPase and a superfamily 3 helicase. The latter is so divergent in the two virophages that primary sequence homology has been largely eradicated. Furthermore, the S3H domain in Sputnik is linked to an N-terminal primase domain, whereas in Mavirus, the S3H domain is not associated with a primase and precedes a C-terminal domain of unknown function (Fig. 4.7). Two additional proteins shared by Mavirus and Sputnik are an adenovirus-type cysteine protease and a short Sputnik protein with a zinc ribbon motif, which is similar to a predicted GIY-YIG endonuclease in Mavirus. The remaining 16 proteins in Sputnik show no sequence similarity to any Mavirus CDSs. Interestingly, although both virophages encode an integrase, these genes belong to different families. Unlike the retroviral integrase found in Mavirus, Sputnik encodes a tyrosine recombinase-family integrase that is similar to archaeal viruses [31]. Whether integrase function is essential for virophage replication remains to be determined. Integrases could ensure long-term survival of the virophage by catalyzing integration of the viral genome into host cell or host virus chromosomes, which would physically link the virophage genome with one of the two organisms that it depends on for replication. The finding that about 80% of Mavirus and Sputnik CDSs appear unrelated argues against a close relationship between the two virophages and suggests that any evolutionary connection between these viruses must be ancient. Alternatively, virophage genomes may experience accelerated mutation rates or exchange genes frequently with cells and other viruses.

4.3.4 Mavirus is related to *Maverick/Polinton* DNA transposons

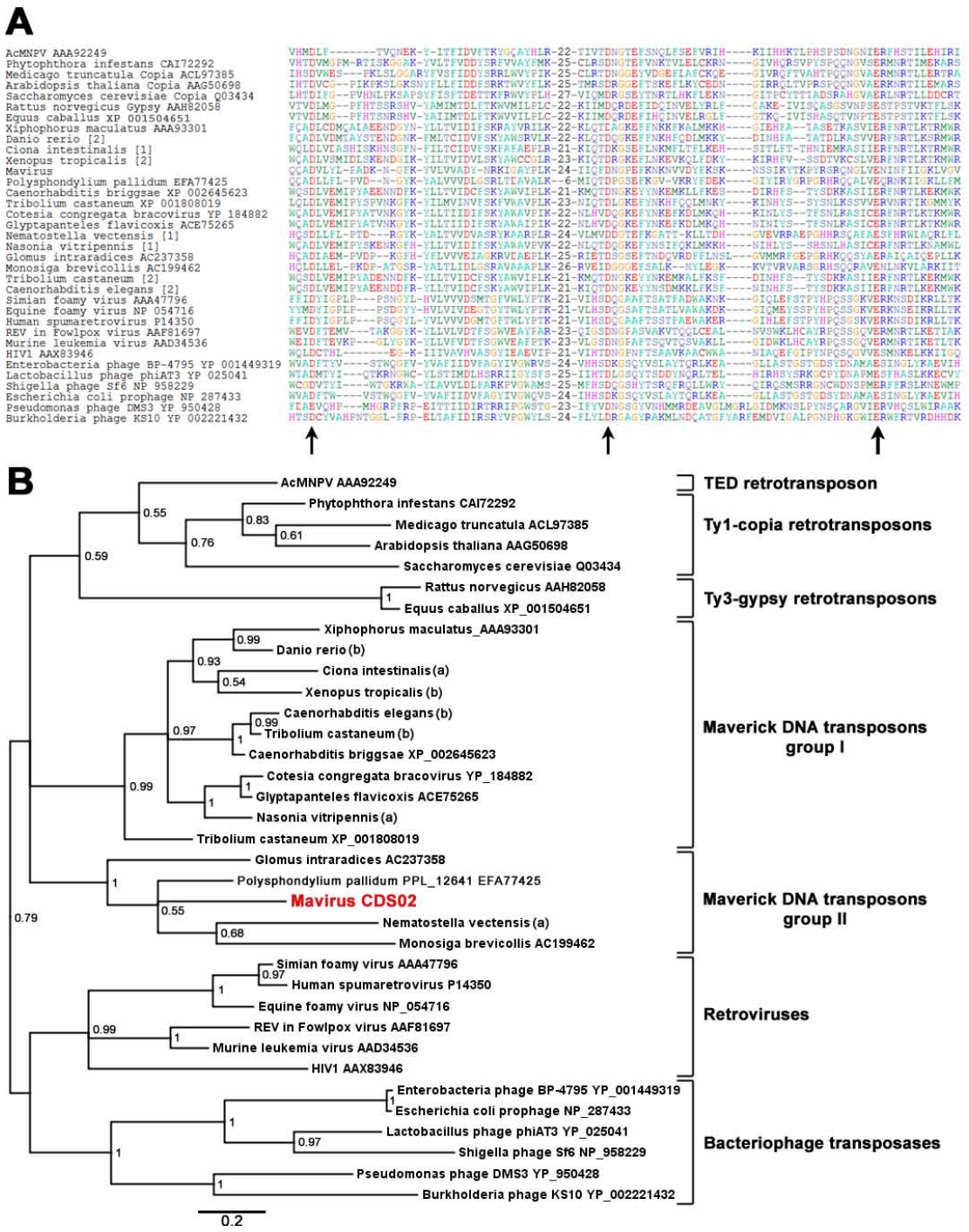
Surprisingly, the Mavirus genome is closely related to a class of eukaryotic DNA transposons. Transposable elements (TEs) can be divided into retrotransposons (class 1 TEs) and DNA transposons (class 2 TEs). DNA transposons are further categorized into cut-and-paste TEs, rolling-circle TEs, and the self-synthesizing *Maverick* or *Polinton* TEs ([258] and references therein). MP TEs are 9-22 kbp long, encode up to 20 proteins, and are found in a wide range of eukaryotes, including protists, fungi, and animals [253, 259, 260]. Their coding capacity varies among species, yet a conserved set of MP genes has been identified [253, 259]. Most MP TEs encode a retroviral integrase, a viral protein-primed DNA polymerase B, an ATPase similar to the FtsK-HerA genome packaging ATPases of dsDNA viruses, and a cysteine protease with homologues in adenoviruses. Another protein termed PY, which is frequently encountered in MP TEs [259], shows structural similarity to the MCP of phycodnaviruses [122]. Because MP TEs genetically resemble viruses of the bacteriophage PRD1-adenovirus lineage [123], a viral origin of these TEs has been proposed [122, 253], although viruses were not known to encode a retroviral integrase in combination with the PRD1-adenoviral genes found in MP TEs.

Homologues of seven Mavirus CDSs are frequently or occasionally encoded by MP TEs (Fig. 4.1A). MV02 belongs to the retroviral integrase family, which contains integrases from retroelements and the eukaryotic c-integrases [248]. Shortly after their discovery, c-integrases were found to be one of the most conserved genes in MP TEs [253, 259, 261]. MP-encoded c-integrases are the closest relatives to Mavirus protein MV02 (Fig. 4.9). Another highly conserved MP gene is the protein-primed DNA polymerase B found in Mavirus MV03 (Fig. 4.3). Two additional genes, the packaging ATPase (MV15) and cysteine protease (MV16), are frequently encountered in MP TEs [253, 259] (Figs 4.5, 4.6). The capsid protein of Mavirus (MV18) is predicted to adopt a secondary structure similar to that of the PY proteins encoded by MP TEs (Fig. 4.8b). Lastly, the superfamily III helicase and MV19 with its C-terminal *B. bacteriovorus* domain are also occasionally encoded by MP TEs.

Not only do Mavirus and MP TEs encode similar genes, but a truncated MP-like element in the genome of the slime mould *Polysphondylium pallidum* shows partial synteny to the Mavirus genome (Fig. 4.10). The syntenic region comprises Mavirus CDSs MV01-MV03 and the rve-INT and DNA polymerase B proteins encoded by *P. pallidum* represent the closest homologues of the respective Mavirus proteins (Figs. 4.3, 4.9, Table 4.3). In addition to the three syntenic genes, the MP-like element in *P. pallidum* includes a predicted cysteine protease and an FtsK-HerA ATPase-like protein (Fig. 4.10). This strongly implies that the *P. pallidum* MP-like element and the Mavirus virophage are genetically related. Furthermore, Mavirus and MP TEs have

Figure 4.9 Phylogenetic position of the Mavirus retroviral integrase.

(**A**) Alignment of integrase core regions of selected retroviruses, retrotransposons, MP TEs, and bacteriophage transposases. Arrows mark the signature conserved amino acids that define the DDE motif [262]. (**B**) Unrooted Bayesian Inference tree based a 91 aa alignment of conserved regions of the rve-integrase core domain. Nodes are labelled with posterior probabilities and GenBank accession numbers are listed where available. Sequences labelled (a) were downloaded from RepBase [252]; sequences labelled (b) were acquired from Pritham *et al.* [253]. AcMNPV, *Autographa californica* multicapsid nucleopolyhedrovirus; REV, reticuloendotheliosis virus.



similar genome lengths (19 kb and 15-20 kb, respectively) and genome structures. MP TEs have terminal-inverted repeats of several hundred nucleotides in length and highly conserved ends that start with 5'-AG and end with CT-3' [259]. Similarly, cutting the circular genome of Mavirus between nucleotides 19,063 and 1 (Fig. 4.1B) would result in a linear DNA molecule with 5'-AG and CT-3' termini, as well as 50 bp terminal-inverted repeats.

4.3.5 Viral transposogenesis of *Maverick/Polinton* DNA transposons

The observed parallels in genome length, genome structure, and gene content to MP TEs in general, as well as the sequence similarity and partial synteny to the MP-like element in *P. pallidum* in particular, imply that the Mavirus virophage and DNA transposons of the MP TEs class have evolved from a common ancestor. The genetic similarities of MP TEs and dsDNA viruses have already spawned various evolutionary hypotheses. It has been suggested that these transposons may have evolved from a linear plasmid [259], whereas others proposed that MP TEs and viruses of the PRD1-adenovirus lineage have evolved from a common ancestral virus [122, 253]. Even the origin of the NCLDVs has been linked to MP transposons [263].

Although one can envision a number of scenarios that would explain the genetic homology between Mavirus and MP TEs, two that would appear to be most likely candidates are that Mavirus was derived from an escaped TE, or that MP TEs arose from the endogenization of an ancient relative of Mavirus. The first scenario is problematic because it requires multiple gene capture events by MP TEs from various sources. For example, the integrase would have to be acquired from a retroelement, whereas the cysteine protease and FtsK-HerA-like ATPase would have originated in a DNA virus. In addition, the first scenario fails to explain the genetic (Figs. 4.5-4.8, Table 4.5) and phenotypic similarities between Mavirus and Sputnik, as well as the dependency of Mavirus on CroV. A more parsimonious explanation is that a Mavirus ancestor (MVA) gave rise to MP TEs. This may have occurred in an early eukaryote susceptible to infection by a large DNA virus (LDNAV) that coevolved with the MVA (Fig. 4.11). MVA replicated by parasitizing the virion factory of LDNAV and interfered with its propagation, much as occurs in the CroV-Mavirus and Mimivirus-Sputnik systems. The interference of MVA with LDNAV replication led to increased survival of the host-cell population, and selection on the host cell to stabilize the relationship with MVA. In turn, the dependence of MVA on LDNAV replication provided strong selection for MVA to be closely associated with the LDNAV host. The fortuitous acquisition of an integrase gene by MVA provided a mechanism for integration into the cellular genome, which may have conferred protection against infection by LDNAV, and selective pressure for MVA to spread throughout the host-cell population. Selection for this viral defence system in early eukaryotes and repeated endogenization of various virophages in

different eukaryotic lineages, combined with accumulation of mutations, intragenomic expansion, and partial or complete loss of the provirophage in some lineages may have led to the current diversity and widespread, albeit patchy, taxonomic distribution of MP DNA transposons (Fig. 4.11) [253, 259].

In addition to their genetic similarities, the evolutionary relationship between Mavirus and MP TEs is supported by several other observations. First, the integrases of Mavirus and MP TEs are closely related, suggesting that an ancestor of the Mavirus integrase could have catalyzed the insertion of DNA into a eukaryotic genome. However, no endogenous forms of Mavirus in uninfected *C. roenbergensis* could be detected by polymerase chain reaction with Mavirus-specific primers. Second, Mavirus, as its ancestors presumably did, uses clathrin-mediated endocytosis to enter the host cell independently of its host virus, as shown by thin-section electron microscopy (see section 5.3.7).This suggests that Mavirus can associate with the host cell even in the absence of a host virus. Third, the presence of capsid-like proteins and other Mavirus-related genes in MP TEs is consistent with their introduction through an ancient Mavirus. Alternatively, there would have to be multiple gene capture events by MP TEs from viral sources, which is much less parsimonious. Finally, Mavirus interferes with CroV propagation and increases the survival of the host-cell population (Table 4.1); hence, eukaryotes that are susceptible to infection by giant viruses will gain a selective advantage if they can associate themselves with virophages. The directionality of the proposed viral transposogenesis is also supported by the shared promoter motif of Mavirus and CroV. It is hard to envision a mechanism that would have led to the *de novo* development from a transposon to the observed host-virus dependence, while it is conceivable that the promoter sequence degenerated after a MVA integrated into the host cell genome.

Figure 4.10 Syntenic regions of Mavirus and an MP-like element from *Polysphondylium pallidum*.

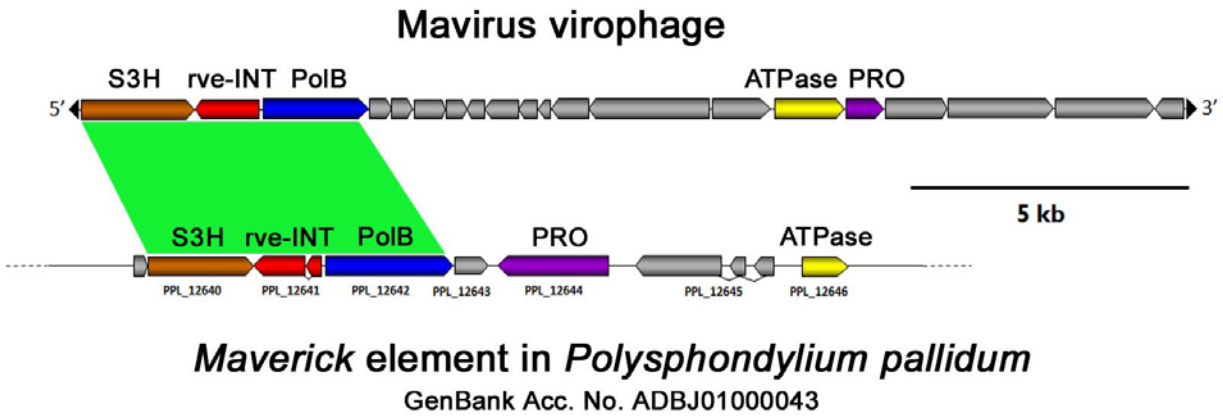
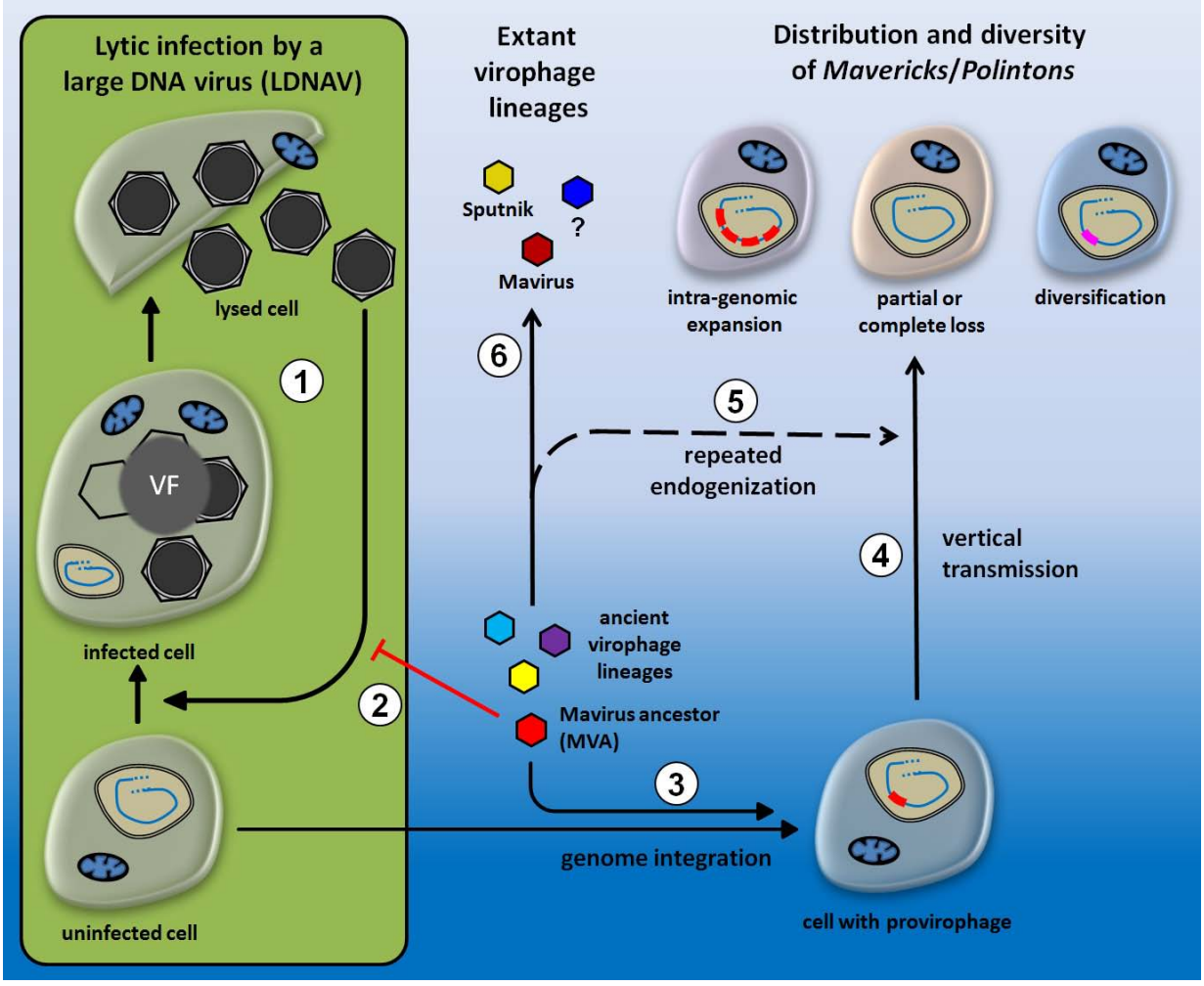


Figure 4.11 Hypothesis for the origin of the *Maverick/Polinton* class of DNA transposons.
(**1**) The infection cycle of a lytic large DNA virus (LDNAV). After virus entry, the virion factory (VF) produces progeny virions that are released during cell lysis. (**2**) A virophage (the Mavirus ancestor, MVA) is dependent on co-infection with the LDNAV and negatively affects its lytic infection cycle. The survival rate of the eukaryotic host population is increased by the MVA. (**3**) Using its integrase, the MVA inserts its genome into the host cell chromosome. (**4**) Over millions of years, the provirophage is vertically transmitted, accumulates mutations, expands horizontally in some eukaryotic lineages (e.g. *Trichomonas vaginalis* [252]), or suffers partial or complete loss in other lineages (e.g. Ascomycota [259]). (**5**) These evolutionary processes are complemented by repeated endogenization of Mavirus-like virophages and result in the observed distribution and diversity of MP TEs. (**6**) Of presumably multiple ancient lineages of virophages, only the one from which Mavirus descended gave rise to the MP TEs. The Sputnik virophage has probably evolved from a different lineage, and many others may remain to be discovered.



4.4 Conclusions

Virophages are a novel group of satellite viruses with only two representatives isolated thus far. Sputnik and Mavirus parasitize the two largest and most complex viruses known, Mimivirus and CroV. Although Sputnik and Mavirus cause similar phenotypes and have genomes of similar size, they differ in their genetic content. It has been proposed that Sputnik could have descended from a mobile element [264]; in contrast, genetic homology implies that an ancestor of the Mavirus virophage gave rise to DNA transposons of the *Maverick/Polinton* class. In this scenario, ancestral forms of Mavirus provided protection to host cells against infection by lytic viruses and were endogenized by the host cells, leading to MP TEs becoming dispersed and fixed within the genomes of many eukaryotic lineages. Although it is not possible to reconstruct these early evolutionary events, the viral transposogenesis hypothesis generates several testable predictions, e.g. whether a virophage can integrate into a eukaryotic genome and whether the resulting provirophage could be induced through superinfection by a host virus. Already, virophages have changed the textbook view of viruses, showing that the largest representatives of the virus world can be parasitized themselves [265]. Through their association with giant viruses and the mosaic nature of their genes, virophages could mediate horizontal gene transfer between giant viruses [31]. The discovery of new virophages will undoubtedly shed further light on the fascinating lifestyle and the evolutionary relationships of these novel superparasites and other forms of mobile genetic elements. In the light of recently described endogenous viruses [266], the discovery of Mavirus reveals a further facet of the intricate genetic interactions between viruses and cellular life.

5 Ultrastructural Characterization of CroV and Mavirus Infection

5.1 Synopsis

Thin-section transmission electron microscopy was used to shed light on ultrastructural aspects of CroV and Mavirus replication in *Cafeteria roenbergensis*. The early stages of CroV infection were characterized by cytoplasmic CroV cores that were membrane-bound and coated with fibres. These cores transformed into early virion factories (VFs), which then grew significantly in size. Mature, capsid-producing CroV VFs developed between 6 h and 12 h post infection (p.i.) and at 18 h p.i., the cytoplasm of infected cells was filled with progeny virions. CroV capsids were assembled at the periphery of the VF and the genome was packaged into preformed capsid structures. Large clusters of ribosome-like aggregates were more frequent in infected cells than in uninfected cells and may support viral protein synthesis. The burst size of CroV was estimated to be around 100 virions per cell.
The Mavirus virophage entered *C. roenbergensis* cells via clathrin-mediated endocytosis and appeared to replicate within the CroV VF, where Mavirus particles were most frequently observed. Progeny virophage particles formed large paracrystalline arrays in the cytoplasm. Putative abnormal CroV capsid structures were often observed in Mavirus-infected cells. The host cell nucleus remained intact until late infection stages and virus particles were never observed within the nucleus, suggesting that CroV and Mavirus replicate exclusively in the cytoplasm.
This study represents the first detailed ultrastructural characterization of the CroV infection cycle and of the intracellular events of a virophage infection. Furthermore, these observations corroborate the classification of Mavirus as a virophage of CroV.

5.2 Materials and methods

5.2.1 Transmission electron microscopy

Exponentially growing *C. roenbergensis* cultures in 250-500 ml modified f/2 enriched sea water medium (see section 2.2.1) (total volume 6 l) were infected with Mavirus-containing CroV lysate at an multiplicity of infection (MOI) of ~0.3 (300 µl of $5x10^7$ CroV ml^{-1}-containing lysate were added to 500 ml of $1x10^5$ cells ml^{-1}). Duplicate aliquots of 380 ml were taken from infected cultures at different time points (0, 3, 6, 12, 18, 24 h post infection) as well as from uninfected

control cultures at 0 h and 24 h p.i.. Samples were centrifuged for 10 min at 5,000 x g in a Sorvall RC-5C centrifuge (GSA rotor, 20°C). The supernatant was discarded, cell pellets were suspended in 0.5 ml f/2 medium and centrifuged for 5 min at 3,000 g in an Eppendorf 5810 centrifuge. Pellets were suspended in 10-15 µl f/2 sea water medium supplemented with 20% (w/v) BSA by stirring with a pipette tip. Cell suspensions were immediately cryo-preserved using a Leica EM HPM100 high pressure freezer. Vitrified samples were freeze-substituted in a Leica AFS system for 2 days at -85°C in a 0.5% glutaraldehyde 0.1% tannic acid solution in acetone, then rinsed 10X in 100% acetone at -85°C, and transferred to 1% osmium tetroxide, 0.1% uranyl acetate in acetone for an additional 2 days at -85°C. The samples were then warmed to -20°C over approximately 10 hours, held at -20°C for 6 hours to facilitate osmication, and then warmed to 4°C over approximately 12 hours. The samples were then rinsed in 100% acetone 3X at room temp, and gradually infiltrated with Spurr's low viscosity embedding media. Samples were polymerized in a 60°C oven overnight. 30-50 nm thin sections and 200-300 nm semi-thin sections for tomography were obtained using a Reichert Ultracut E ultramicrotome and a Drukker diamond knife. The 30-50 nm sections were stained for 12 min in 2% aqueous uranyl acetate and 6 min in Reynold's lead citrate. Staining times for the 200-300 nm sections were increased to 30 min and 15 min respectively. Image data were recorded on an FEI Tecnai G2 20 TWIN transmission electron microscope at 80 kV for the 30-50 nm sections and 200 kV for the 200-300 nm sections. Thin sections in figures of this chapter were 30-50 nm thick unless stated otherwise. Adobe Photoshop v7.0 was used to compile the TEM images. Adjustments to contrast and brightness levels were applied equally to all parts of the image.

5.2.2 Estimation of burst size

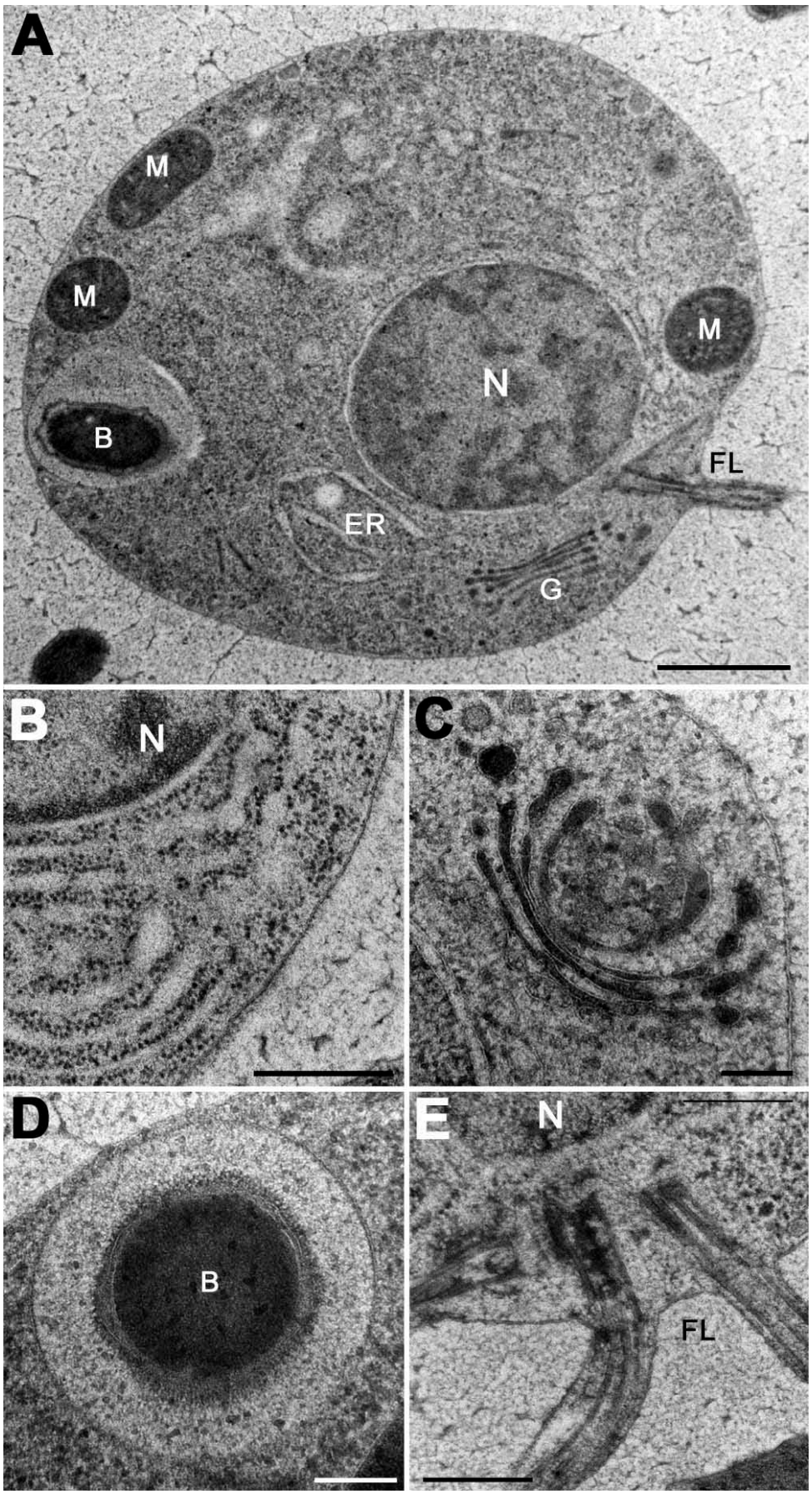
The burst size was estimated by counting the number of virus particles in 300 nm thin sections of visibly infected cells at 18 h p.i. and 24 h p.i. and extrapolating the counts to the entire cell volume. The cell volume was estimated by assuming that the shape of *Cafeteria* cells can be approximated by a sphere. For each visibly infected cell, the number of virus particles was counted and cell dimensions were measured at the lowest and largest diameters, resulting in an average cell radius. The thin sections had a thickness of 300 nm, but a correction for calculating the volume of the section was necessary in order to account for virus particles that were only partially visible. Assuming a maximum particle size of 300 nm for CroV, half a virion diameter was added to the top and bottom of the slice, which resulted in a corrected slice thickness of 600 nm.

5.3 Results and discussion

5.3.1 Ultrastructure of *Cafeteria roenbergensis*

The phagotrophic nanoflagellate *C. roenbergensis* Fenchel & Patterson is a single-celled eukaryote in the kingdom Chromista (Stramenopiles) [136]. The 3-6 µm diameter cells are kidney shaped and possess two flagella, one anterior and one posterior flagellum. The anterior flagellum is covered bilaterally by tripartite tubular hairs. The flagella emerge from the ventral side of the cell and are anchored by three microtubular roots that also form part of the feeding apparatus [267]. During feeding, the posterior flagellum is attached to a substrate, while the anterior flagellum beats to create a current that directs potential prey items towards the cell. In swimming mode, the anterior flagellum beats and pulls the cell forward. Figure 5.1 shows the fine structure of a *C. roenbergensis* strain E4-10 cell. The ovoid nucleus is usually located in the centre and is associated with one of the two basal bodies at the flagellar base (Fig. 5.1A,E). *C. roenbergensis* contains three or more mitochondria with tubular cristae, as is typical for Stramenopiles. The Golgi apparatus is situated at the ventral side (defined by the flagellar apparatus) of the cell (Fig. 5.1A,C). The posterior end often contains phagocytic vacuoles containing ingested bacterial prey (Fig. 5.1A,D). A rather unusual feature of *Cafeteria* is the presence of extrusomes, which are bottle-shaped vesicles 100-200 nm in diameter that deliver their electron-dense contents to the cell exterior [136, 267]. However, extrusomes were rarely observed in *C. roenbergensis* strain E4-10.

Figure 5.1 Ultrastructure of Cafeteria roenbergensis strain E4-10 (next page).
(**A**) Thin section of a *C. roenbergensis* cell, strain E4-10. This cell features a large nucleus (N) next to the flagellar apparatus (FL) and the endoplasmic reticulum (ER), a Golgi body (G), three mitochondria (M) and a food vacuole with a phagocytosed bacterium (B). Scale bar = 1 µm. (**B**) Rough endoplasmic reticulum. Scale bar = 500 nm. (**C**) Golgi body with the *cis* face on the lower left and the *trans* face on the upper right. Scale bar = 200 nm. (**D**) Phagocytic vacuole containing an ingested bacterium. Scale bar = 200 nm. (**E**) The flagellar apparatus of *C. roenbergensis*. Scale bar = 500 nm.


A
M
M
M
N
B
FL
ER
G
B
N
C
D
B
E
N
FL

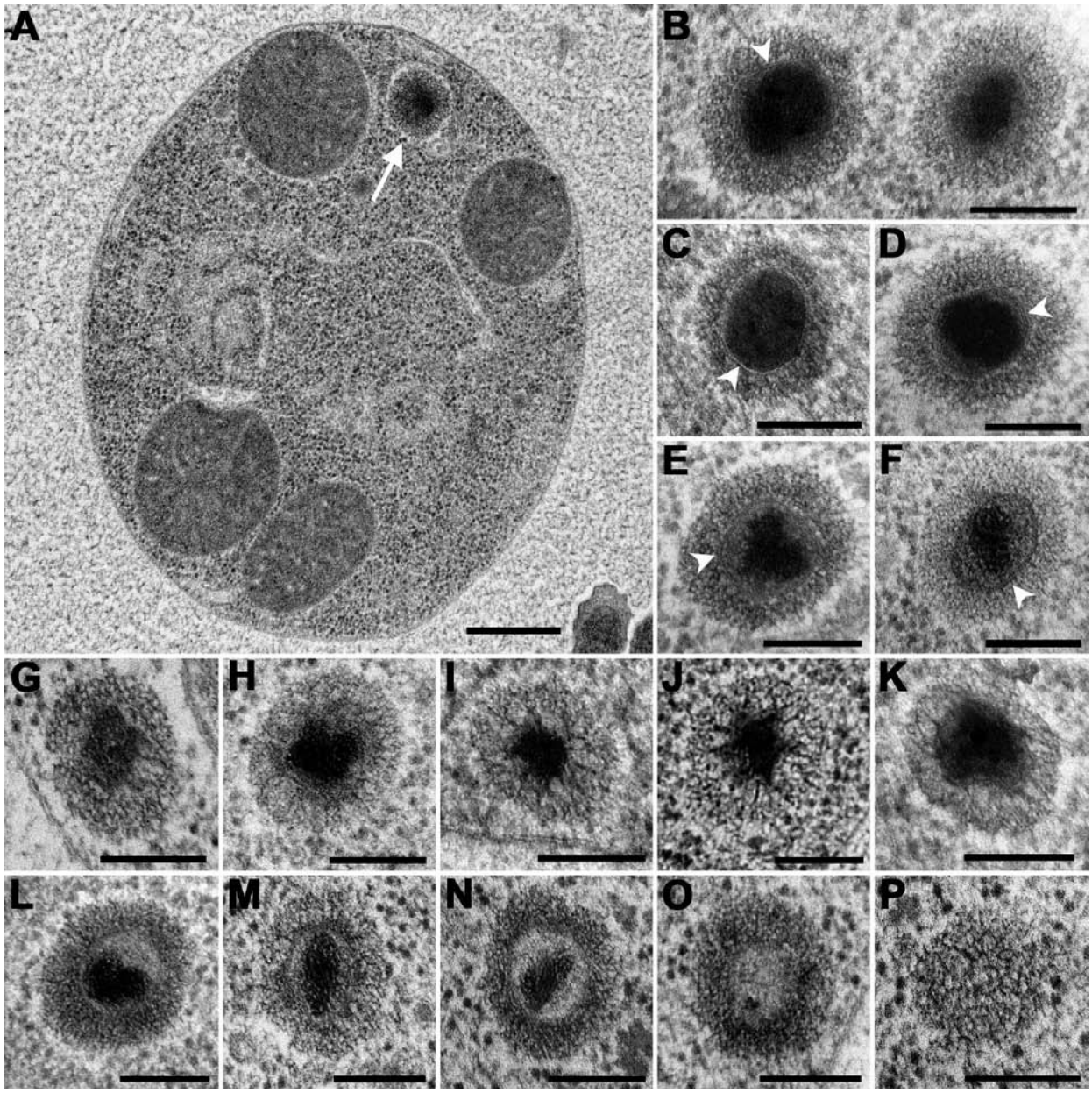
5.3.2 CroV cell entry and uncoating

The ultrastructural phenotype of *C. roenbergensis* cells infected with CroV and Mavirus was studied by thin-section transmission electron microscopy (TEM). Exponentially growing cultures were infected with a lysate (labelled "25.10.07 B") containing approximately $5x10^7$ CroV ml^{-1} and an unknown concentration of Mavirus. The low CroV MOI of ~0.3 was chosen because higher MOIs had frequently led to unproductive infections and premature cell death. Whether this effect was due to the presence of the virophage or the "lysis from without" phenomenon during which too many penetration events can lead to severe damage of the cytoplasma membrane [209, 210], is not known. Probably as a direct consequence of the low MOI used, no penetration and virus entry events of CroV were observed. A possible endocytic mode of entry for CroV is discussed in section 5.3.3.

The first sign of CroV in infected cells was the occurrence of round cytoplasmic objects that had an electron-dense core with a diameter of 160 nm (± 28 nm, n=21). As shown in Figure 5.2, the cores were membrane-bound and surrounded by a ~70 nm thick layer of fibres. These structures were observed in all infected samples, starting at 0 h p.i., but were not present in uninfected cells. Furthermore, the electron-dense centre corresponded in its dimensions to the core of mature CroV virions (170 nm ± 17 nm, n=43). It can therefore be assumed that these particles represent CroV cores. None of the cores were associated with a viral capsid, suggesting either that the capsid did not enter the cell during penetration and core delivery, or that the capsids are unstable after uncoating. Some cells displayed more than one CroV core (Fig. 5.2B); superinfection by several CroV virions is therefore possible. This is similar to findings from Mimivirus, where multiple cores were found to seed independent virion factories (VFs), which may fuse during late infection [54, 67]. The chemical nature of the fibres that surrounded the CroV cores is unknown; however, given that CroV packages a large arsenal of transcription proteins and is likely to transcribe its early genes inside the core upon penetration (see section 3.3.3), these fibres may consist of viral mRNA exiting the core through membrane pores. The electron-dense material inside the CroV cores, presumably the viral genome, sometimes displayed a uniform density (e.g. Fig. 5.2B,C), but most of the cores seemed only partially filled (Fig. 5.2D-N). In some cases, only the surrounding layer of fibrils was still visible (Fig. 5.2O), which may be an intermediate stage in the transition to early VFs (Fig. 5.2P).

Figure 5.2 Cytoplasmic CroV cores during early infection.

(**A**) Cross section of a *C. roenbergensis* cell featuring a fibre-coated CroV core (arrow). Scale bar = 500 nm. (**B-P**) Various morphologies of CroV cores. The core membrane, from which the dense fibres seem to protrude, was visible in some cases (marked by arrow heads). Dark dots surrounding the viral cores are free ribosomes. The transition from viral core to early VF is shown in (N-P), as the electron-dense centre material disappeared and the surrounding layer of fibres adopted a ball-like structure. These structures were not observed in uninfected cells. Scale bars = 200 nm.



5.3.3 The CroV virion factory

The virion factory (VF, also called virus factory) is a specialized cytoplasmic compartment that is generated by viruses to replicate their genomes and assemble their virus particles [268]. VFs often recruit ribosomes and mitochondria in their vicinity to supply the VF with proteins, ATP, and other essential substrates [82]. Large DNA viruses such as the NCLDVs are known to build the largest and most spectacular VFs [67, 268, 269]. During viral infection, the VF expands significantly until it eventually fills most of the cell volume. In CroV-infected cells, VFs at different maturation stages could be observed (Fig. 5.3). Early VFs were sometimes difficult to identify, especially at lower magnification, due to the lack of contrasting features such as electron-dense virions surrounding them. However, the filamentous ultrastructure of VFs at higher magnification was very similar to the outer fibrous layer of early CroV cores and was clearly distinguishable from the surrounding cytoplasm, which had a more granular appearance due to the presence of free ribosomes (Fig. 5.3A). Only VFs that had grown to approximately 1 µm in diameter or larger were found to produce capsid structures and progeny virions (Fig. 5.3C,D). Virus-like particles (VLPs) were first observed in 12 h p.i. samples, but not in 6 h p.i. samples. This suggests that CroV assembly starts between 6 h and 12 h p.i. and confirms results from microarray and gene promoter analyses, which revealed that CDSs encoding structural proteins were associated with a later promoter element and were first transcribed at 6 h p.i. (see sections 2.3.10 and 2.3.11). Higher magnification of mature VFs revealed that capsid structures were assembled at the periphery of the VF, where they were prefabricated prior to genome packaging. Examples of CroV capsids at different assembly stages are shown in Fig. 5.4. Capsid fabrication starts with the assembly of one particular face or vertex of the icosahedral particle (arrow heads in Fig. 5.4A), which moves away from the VF as polymerization progresses (arrows in Fig. 5.4A). Nascent capsids appeared to be lined with a second, inner layer (arrow head in Fig. 5.4B), possibly a lipid membrane that could be attached to the capsid structure via transmembrane protein interaction (see section 3.3.2). Structurally complete, but empty capsids were frequently observed along the VF periphery (arrows in Figs. 5.4C and 5.5A), although a few cases were documented where intact, yet empty or only partially filled capsids were detached from the VF and enclosed in a vesicle (Fig. 5.5B,C). It is possible that these cases represent endocytic vesicles containing remnants of the virion which infected the cell. This would imply that CroV enters *C. roenbergensis* cells via endocytosis and that the viral core is released into the cytoplasm with the capsid structure remaining in the endocytic vesicle. Because the cross-sections in Fig. 5.5B and C show structurally intact capsids, it is likely that the capsid opening occurs at one particular site, perhaps at a unique vertex similar to the stargate opening in Mimivirus [66].

Figure 5.3 Different stages during expansion of the CroV virion factory.

(**A**) Two early VFs are marked by arrows. The inset shows a higher magnification of the boxed area. Scale bars = 500 nm (main image), 200 nm (inset). (**B**) Cell containing a medium-sized VF (arrow) showing no signs of capsid assembly. Scale bar = 500 nm. (**C**) The VF in this cell has already started to produce virions (arrowhead), but most of the cytoplasm is still unaffected by viral infection. Scale bar = 500 nm. (**D**) Cell displaying a large VF and several mature virions (arrowheads) in the cytoplasm. Scale bar = 500 nm. Symbols: FL, flagellar apparatus; M, mitochondrion; RA, ribosome aggregates; VF, virion factory.

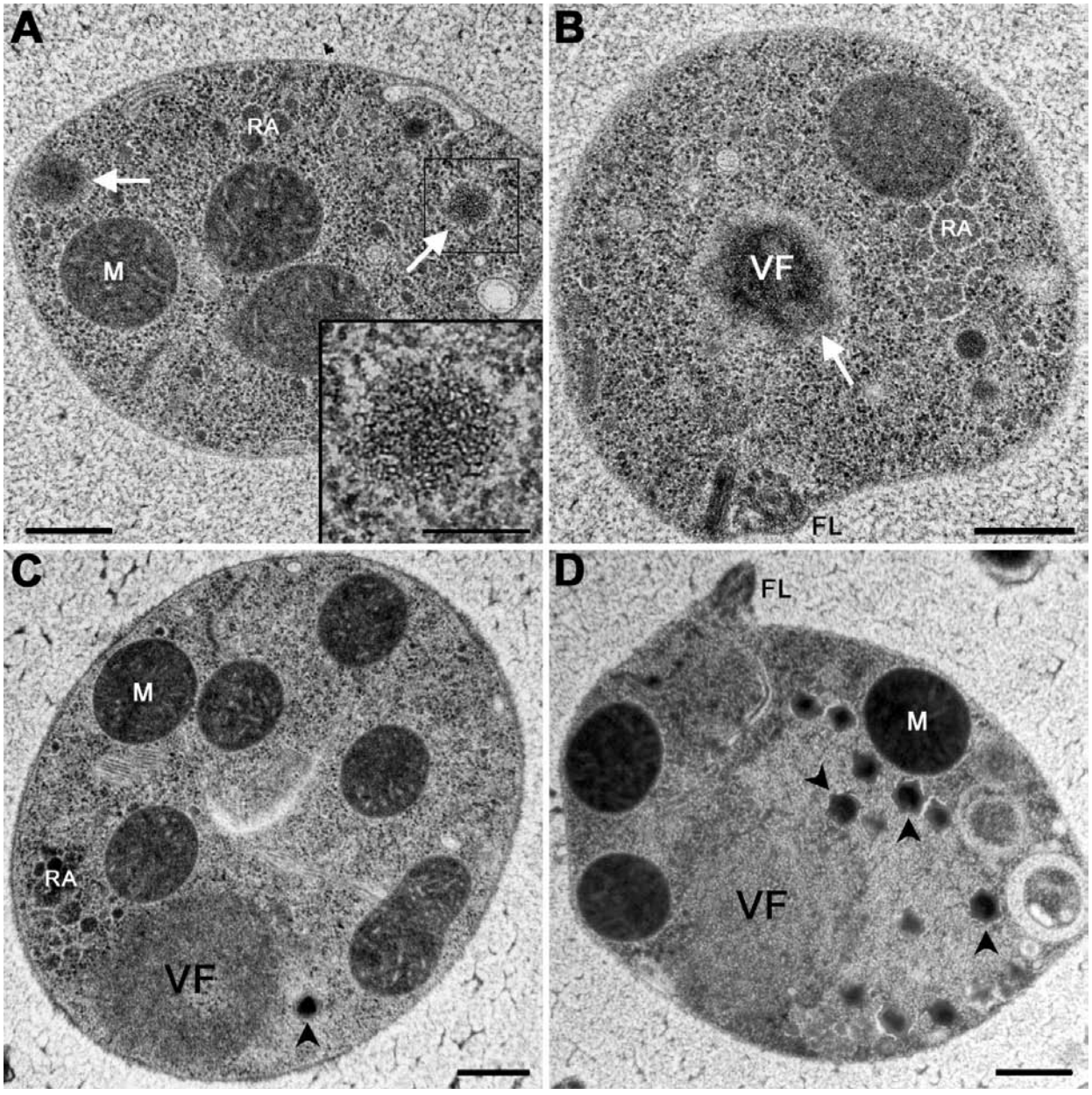


Figure 5.4 Ultrastructural aspects of CroV capsid assembly.

(**A**) Early (arrowheads) and intermediate (arrows) stages of CroV capsid assembly. (**B**) A partially assembled capsid displaying a second layer beneath the outer capsid layer (arrowhead). (**C**) Partially assembled capsid (arrowhead) and intact empty capsid (arrow). All images are oriented so that the VF is located at the bottom and the cytoplasmic side at the top. All scale bars = 200 nm.

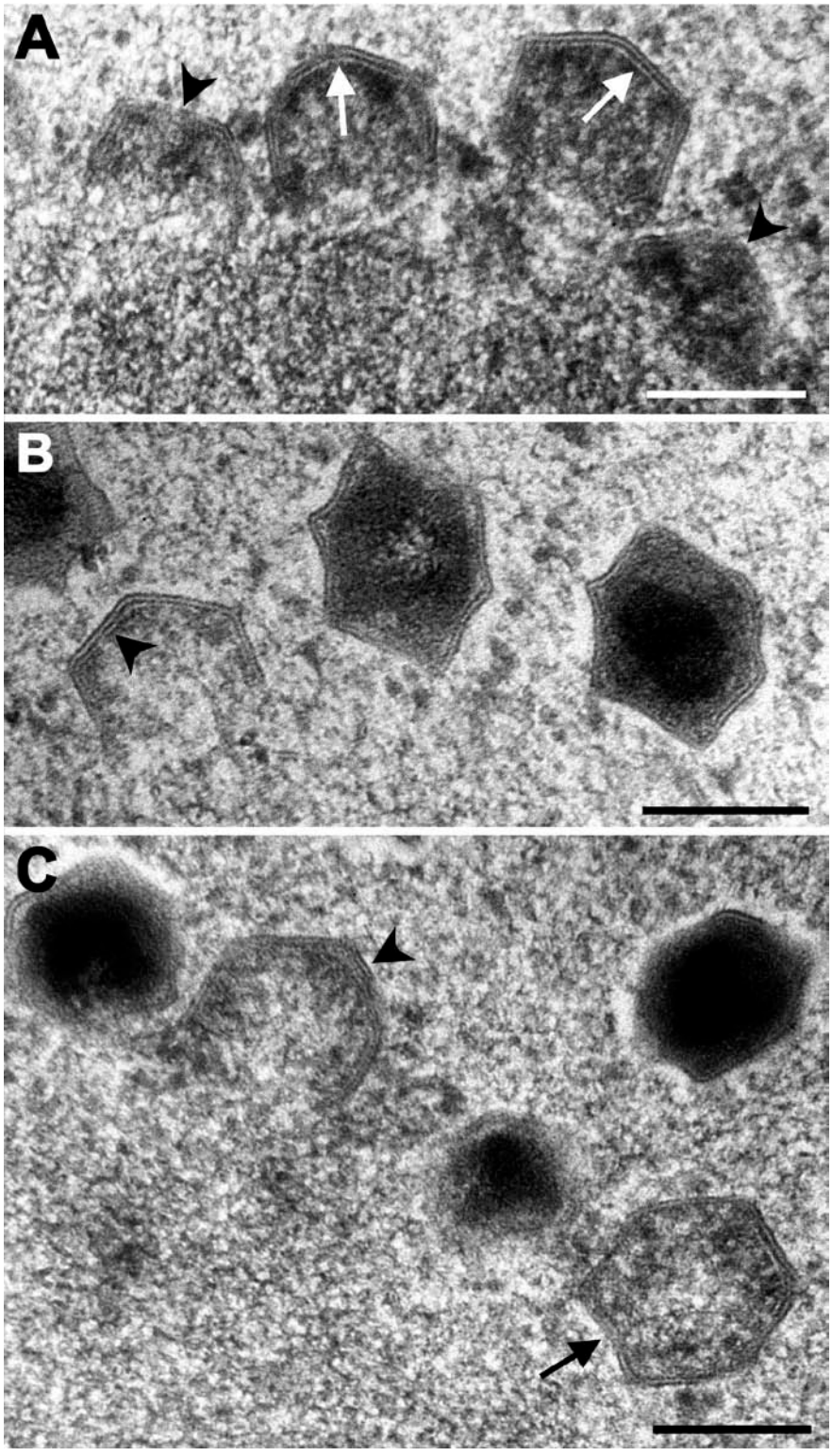
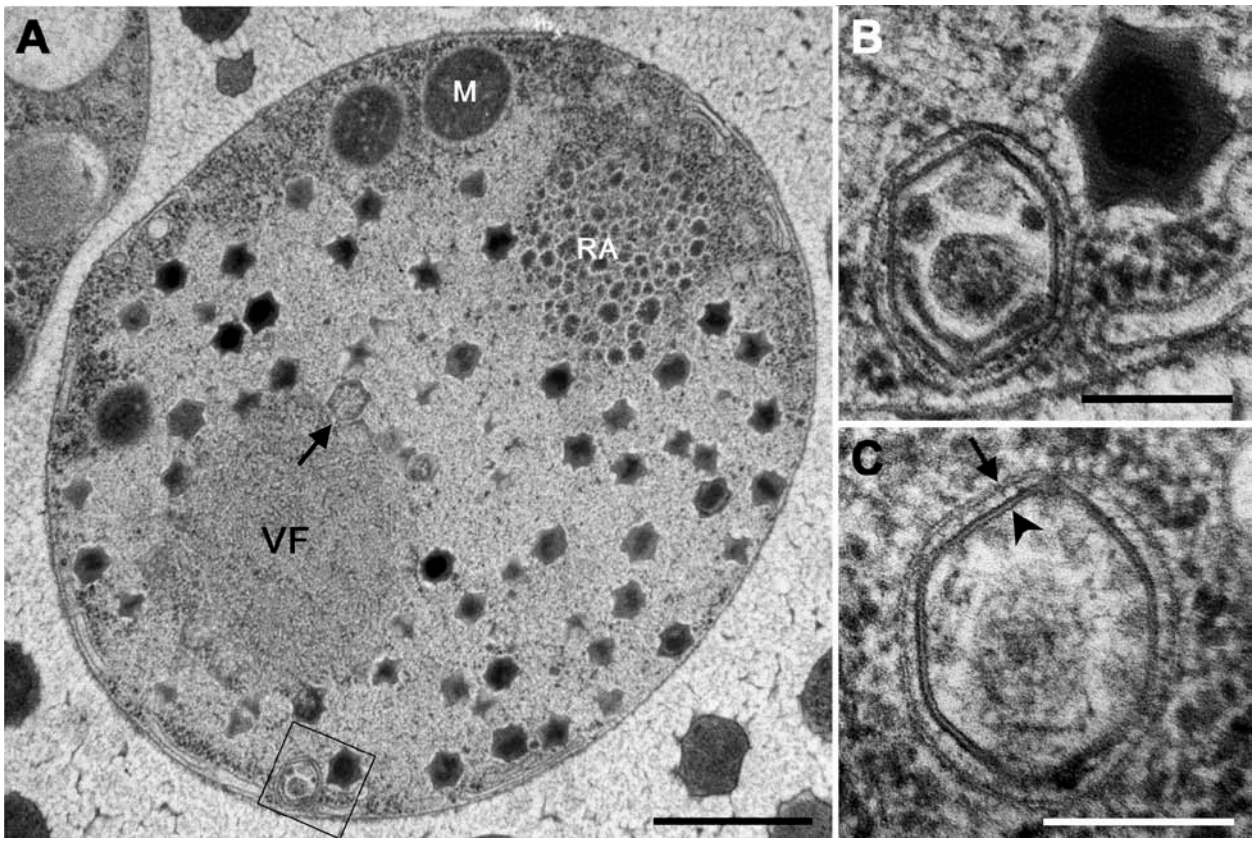


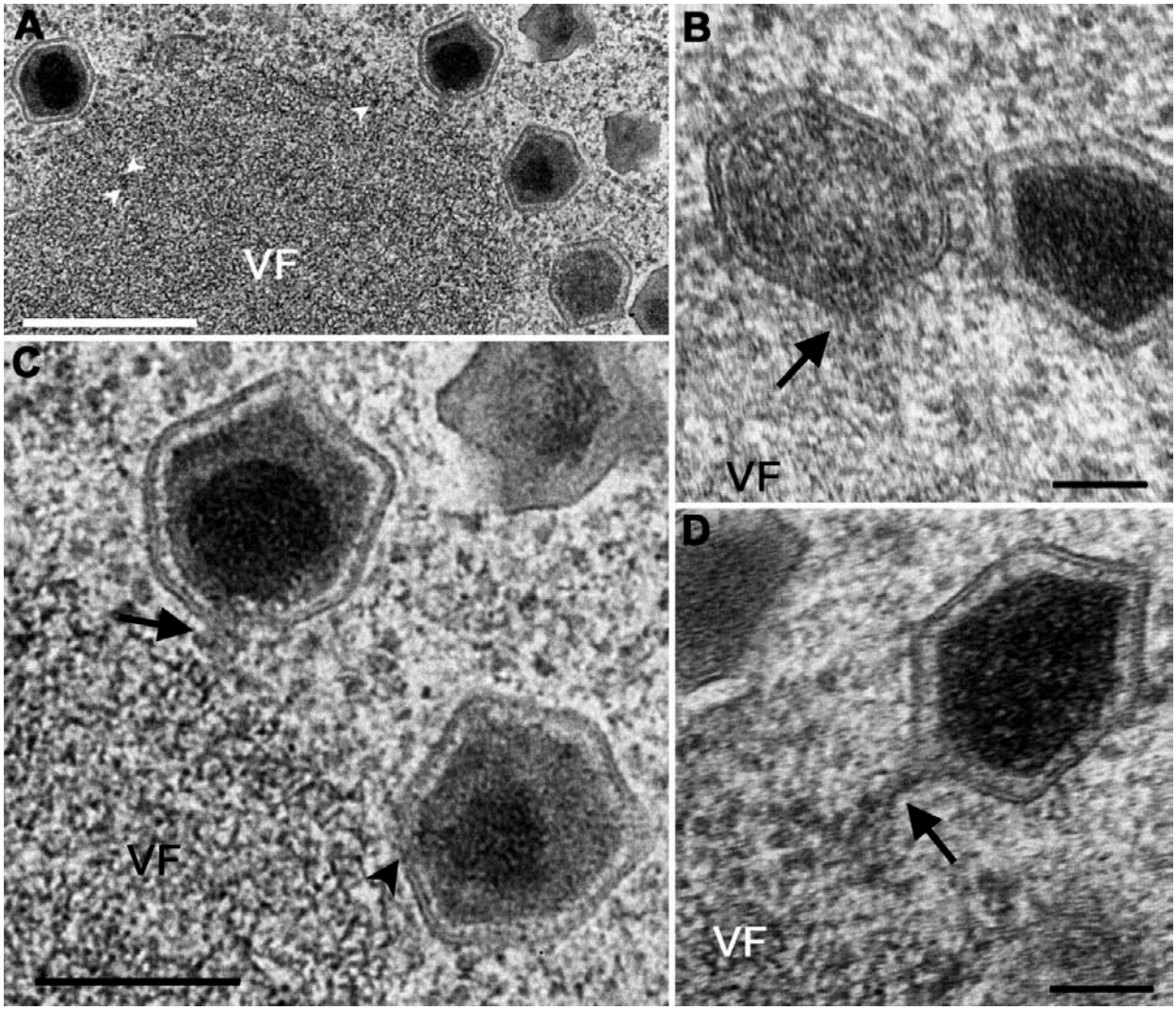
Figure 5.5 Mature CroV virion factory and vesicles bearing empty capsids.
(**A**) Visibly infected cell with large VF and numerous mature CroV virions. An empty but otherwise complete capsid is marked by an arrow. Scale bar = 1 µm. (**B**) Higher magnification of the boxed area in (A), showing a vesicle-enclosed capsid structure next to a slightly deformed mature CroV virion. Scale bar = 200 nm. (**C**) Membrane-bound empty capsid structure from a different cell. The arrow marks the lipid envelope, the arrowhead marks the capsid layer. Scale bar = 200 nm. Symbols: M, mitochondrion; RA, ribosome aggregates; VF, virion factory.



Packaging of the viral genome occurs after capsid assembly, as electron-dense matter was only found in visually intact capsids. Figure 5.6A shows VF-associated CroV particles with heterogeneous core densities, indicating different stages of genome packaging. Capsid filling may occur in two distinct stages, because two distinguishable levels of core densities were observed. The immature particle on the left in Fig. 5.6B (arrow) represents the first stage, exhibiting a uniform core density that was higher than that of the surrounding cytoplasm, but lower than that of the mature virion on the right. The second stage is shown in Fig. 5.6C, where highly dense material is injected on top of the lesser dense material already present in the capsid. It is unclear whether DNA and proteins are packaged simultaneously or successively.

Figure 5.6 Ultrastructural details of CroV particle filling.

(**A**) CroV particles attached to the VF displaying various stages of genome packaging. Single ribosomes within the VF are marked by arrowheads. Scale bar = 500 nm. (**B**) Two virions illustrating the two successive stages of capsid filling. The particle on the left has a lower core density than the mature virion on the right. The capsid opening through which the lesser dense material seems to be injected is marked by an arrow. Scale bar = 100 nm. (**C**) Higher magnification of two CroV particles from (A). The capsids are filled through openings that are located at a vertex (arrow) or at a capsid face (arrowhead). Scale bar = 200 nm. (**D**) A completely packed virion is still attached to the VF (arrow). Scale bar = 100 nm.



The final stage of virion production is shown in Fig. 5.6D, which displays a particle with the uniformly dark core typical of mature virions; however, this capsid was still attached to the VF by a thin thread (arrow). Most capsids were filled through a vertex of the capsid (arrows in Fig. 5.6B-D), only rarely was the opening located at the face of a capsid (arrowhead in Fig. 5.6C).

5.3.4 The CroV virion and cell lysis

Mature intracellular CroV virions were 220-280 nm in diameter and pentagonal or hexagonal in cross-section, confirming previous assumptions of icosahedral symmetry [101]. It is known that virions of large DNA viruses tend to shrink by 16-24% during sample preparation for thin sectioning [36]. Assuming an average diameter of 250 nm for CroV capsids in TEM thin sections and an average shrinkage of 20%, the true diameter of the CroV virion may be ~300 nm. Whereas partially assembled or empty capsids appeared mostly intact and symmetrical, mature virions with electron-dense cores were often deformed and squished (e.g. Fig. 5.5B, dense particle on the right). This might be an artefact of sample preparation, as pressurized mature virions are probably more susceptible to osmotic damage than empty capsids. A selection of intact mature cytoplasmic virions is presented in Fig. 5.7. The outer capsid layer was clearly discernible in all particles (marked by arrows in Fig. 5.7C-E), whereas lipid layers were not always visible (e.g. Fig. 5.7A). The CroV capsid probably contains two internal lipid membranes, one directly beneath the capsid layer and a second one surrounding the genome-containing dense core (arrowheads in Fig. 5.7C-E). The presence of lipids was also suggested by a loss of infectivity after chloroform treatment, as reported by Garza and Suttle [101]. Some TEM images suggest that CroV virions may have two unique, opposing vertices. These vertices (marked by asterisks in Fig. 5.7C,E) were slightly elongated compared to other vertices, which resulted in an increased intermembrane space underneath these vertices. Unique vertex structures have been described in PBCV-1 and Mimivirus, and are likely to facilitate genome release in these viruses [62, 270]. However, cryo-electron microscopy of CroV particles would be a better suited method to investigate such a detailed ultrastructural feature.

Newly synthesized virions accumulated in the cytoplasm until cells lysed. No events of virions budding from the host plasma membrane were observed. Some cells were almost entirely filled with virus particles, such as the 24 h p.i. cell in Fig. 5.8A. The plasma membrane of this cell may have just started to lyse (arrow). A lysed cell containing only a few remaining mitochondria and VLPs is shown in Fig. 5.8B. Cell lysis events were rare in TEM thins sections, as these cells were probably not very stable and may have disintegrated during sample preparation.

Figure 5.7 Mature intracellular CroV virions.

(**A-F**) Capsid layers are marked by arrows, lipid layers by arrow heads. Asterisks indicate possible unique vertices. All scale bars = 100 nm.

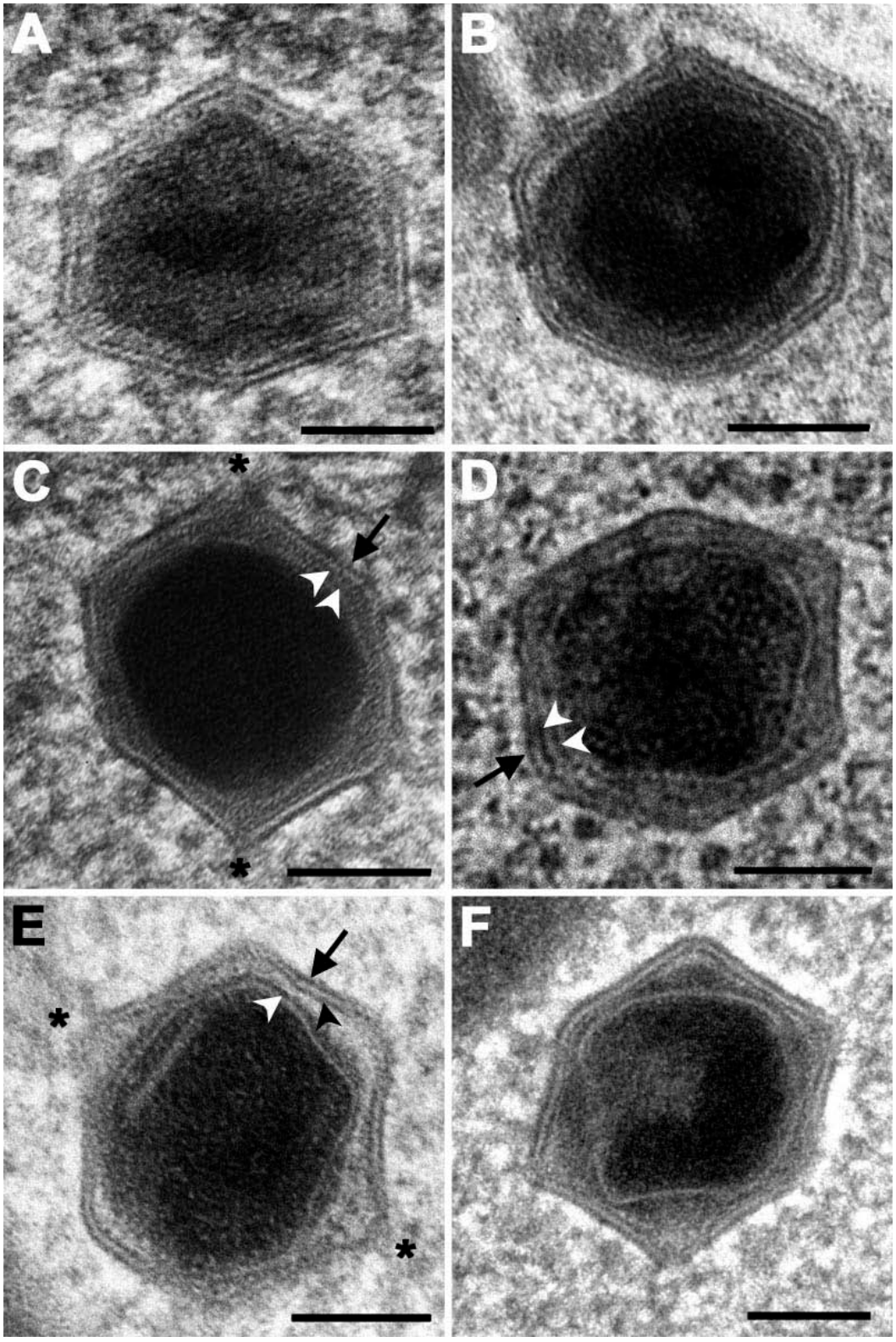
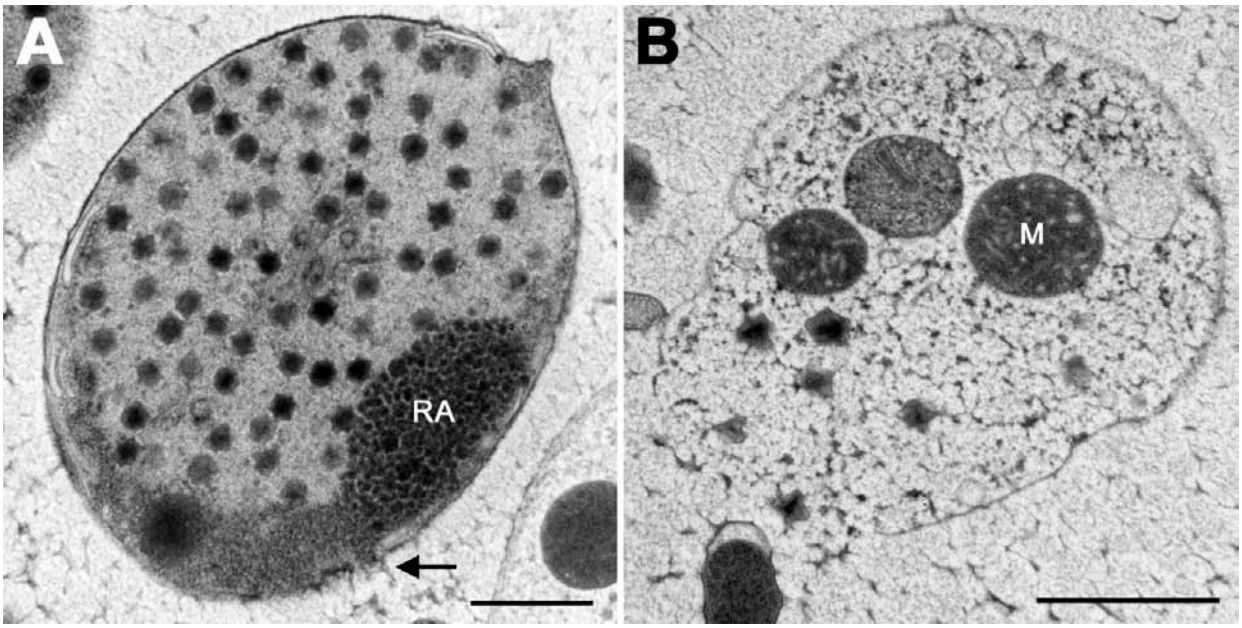


Figure 5.8 Late stages of CroV infection.

(**A**) 300 nm thin section of a cell filled with progeny CroV particles and showing signs of beginning lysis (arrow). (**B**) Lysed *C. roenbergensis* cell. All scale bars = 1 µm. Symbols: M, mitochondrion; RA, ribosome aggregates.



5.3.5 CroV infection may be supported by ribosome aggregates

Viruses have evolved many ways to orchestrate the host translation system for the purpose of viral protein synthesis [162]. Large DNA viruses such as the NCLDVs sometimes encode their own translation initiation and elongation factors, which may be used to gain additional control over host translation (see section 2.3.2). Poxviruses and African swine fever virus (ASFV) build large VFs in the cytoplasm, which are surrounded by mitochondria for ATP supply and often contain membranous material, ribosomes and other translation machinery to facilitate viral protein synthesis and virion assembly. Experiments in vaccinia virus have shown that host translation factors and ribosomes are relocated from the cytoplasm to the VF [269]. Similar reports from ASFV indicate that host translation factors are actively redistributed by the virus inside and near the VF [271]. Viruses therefore actively recruit the host translation machinery to the site of viral mRNA production, rather than transporting transcripts to the ribosomes and the resulting proteins back to the VF.

The CroV VF seemed largely exclusive to ribosomes (note the few ribosome particles marked by arrowheads in Fig. 5.6A), but the periphery of VFs frequently contained free ribosomes (e.g.

Figure 5.9 Ribosome aggregates in *C. roenbergensis*.
(**A**) Cell displaying a fibrous CroV core (arrow), mitochondria (M) and a large cluster of ribosome aggregates (boxed area). Individual ribosome aggregates are also present (arrowheads). Scale bar = 1 µm. (**B**) Higher magnification of the boxed area in (A). The aggregates are composed of granules that resemble free ribosomes. Scale bar = 200 nm. (**C**) Ribosome aggregates from a different cell. This particular cluster had a diameter of ~1.5 µm. Scale bar = 200 nm (**D**) Close-up view of ribosome aggregates from an uninfected cell. Scale bar = 200 nm.

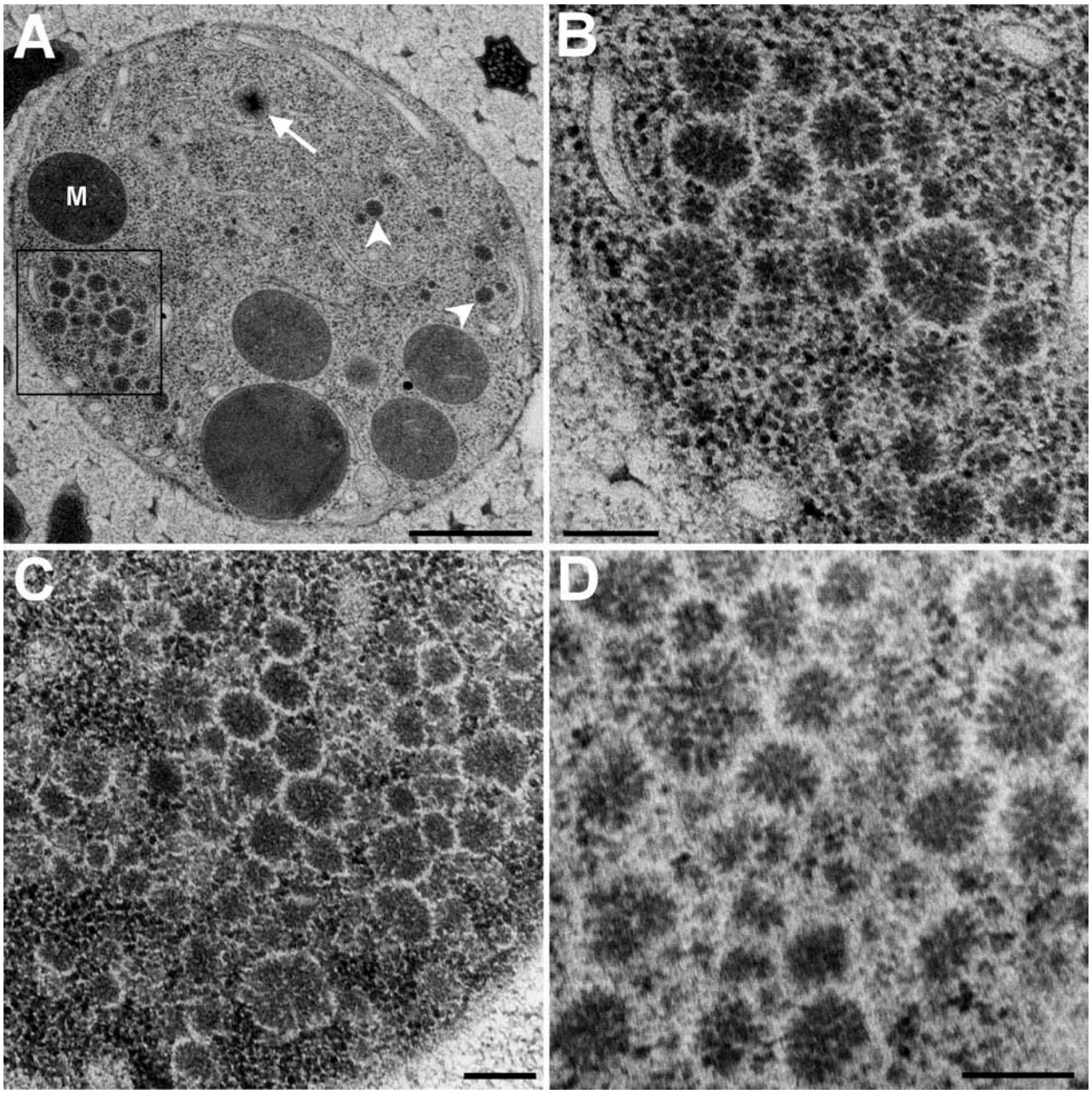


Fig. 5.6A). Interestingly, CroV-infected cells often displayed large cytoplasmic structures that resembled rosette-like aggregates of non-ER-associated ribosomes (Fig. 5.9). These aggregates had a diameter that varied from 60 nm to 250 nm and consisted of 20-25 nm large

granules, which corresponded in size and appearance to free ribosomes. It can therefore be assumed that these structures are composed of ribosomes and I will henceforth refer to them as ribosome aggregates. Although individual rosettes of ribosome aggregates were sometimes visible (arrowheads in Fig. 5.9A), they mostly occurred clustered in specific cellular locations (boxed area in Fig. 5.9A). No literature has been found that described similar structures. It is possible that cryofixation may preserve these structures better than glutaraldehyde fixation; the latter method was used for previous studies of *Cafeteria* fine structure [136]. Whereas only few uninfected cells displayed ribosome aggregates (Fig. 5.9D), they were generally more frequent and also larger in infected cells (Fig. 5.9B,C). The largest clusters of ribosome aggregates were observed during late infection, when cells contained mature, virion-producing VFs (Figs. 5.3C, 5.5A). In cells at very late infection stages, these clusters were often the only remaining host structures (Fig. 5.8 A). Although further experiments will be necessary to elucidate this novel subcellular structure, the preliminary TEM data suggests that large clusters of ribosomes may serve as production sites for viral proteins outside the VF.

5.3.6 CroV burst size

The burst size can be estimated as the average number of viral particles in visibly infected cells [272]. In thin sections, these counts have to be extrapolated to the entire cell volume, which poses additional problems if the cell shape is irregular or if organelles remain in the cell, which reduces the volume available to virions. In order to estimate the CroV burst size in *C. roenbergensis*, 22 images of visibly infected cells at 18 h p.i. and 24 h p.i. were analyzed. The number of VLPs in these 300 nm-thick sections ranged from 14 to 62 with an average of 30.3 ± 11.7. Extrapolated to the entire cell volume, which was approximated to be spherical, the burst size of CroV in these cells ranged from 49 to 311, with an average of 135 ± 62 particles per cell. Although this number is only a rough approximation, it shows clearly that the burst size of this giant virus is on the order of a hundred to a few hundred particles, rather than thousands. For comparison with other large DNA viruses, PBCV-1 has a burst size of 200-250 infectious units per cell [273], *Ostreococcus tauri* virus 5 has a burst size of ~25 [274], *Heterosigma akashiwo* virus 01 has a burst size of ~770 [275], and the burst size of Mimivirus is estimated to be larger than 300 particles per cell [59].

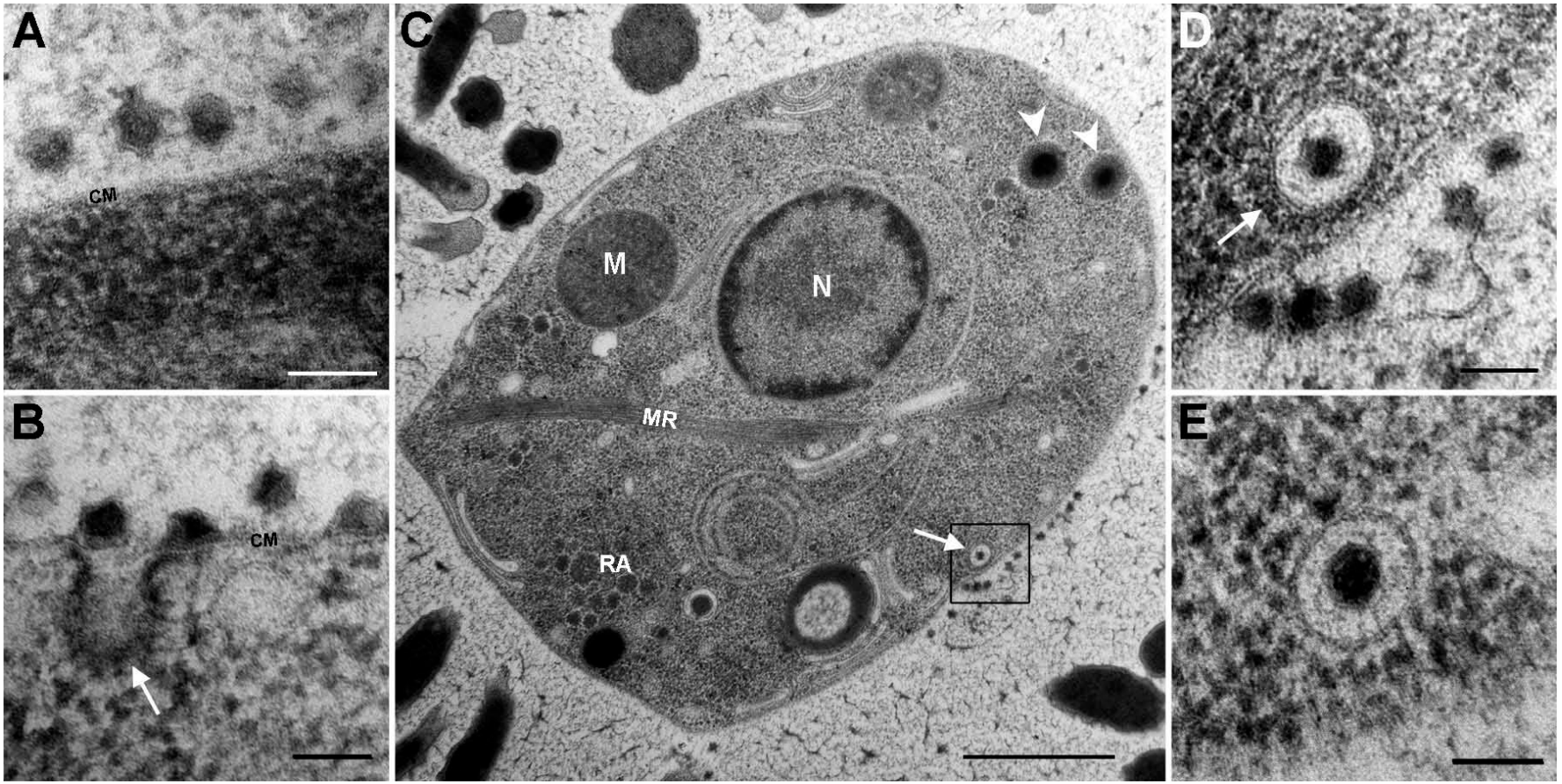
5.3.7 Clathrin-mediated endocytosis of Mavirus

Virophages are a novel type of satellite viruses that replicate within the VF of a giant helper virus and negatively affect the replication of their helper virus, thereby increasing survival of their host cells [31, 105] (see section 1.3). A virophage parasitizing CroV was discovered and

named Mavirus (see chapter 4). Mavirus is only the second described virophage; hence, many aspects of virophage biology are still unknown, including the mechanism of entry. Being dependent upon co-infection by a giant virus for their replication, virophages could also rely on their host virus for gaining entry into the host cell. For Sputnik, it has been hypothesized that interaction with the fibrils of Mamavirus could allow the virophage to "piggyback" its way into the amoeba [254]. This is not the case for Mavirus, as ~60 nm large virions were observed at different stages of endocytosis without apparent involvement of CroV (Fig. 5.10). In 24 h p.i. thin sections, virophage particles were frequently observed attached to the cell surface, usually at a distance of ~30 nm from the plasma membrane (Fig. 5.10A). This implies that the cell surface, the virion surface, or both may be coated with fibres which could be involved in virophage-cell interaction. Invagination of the plasma membrane at sites of attached virophages was observed, as shown in Fig. 5.10B. The inner surface of the invagination appeared thicker and denser than a normal cytoplasma membrane (arrow). These structures had striking resemblance to clathrin-coated pits, which are formed during clathrin-mediated endocytosis (CME) [276, 277]. CME is one of several endocytic pathways by which viruses gain entry into cells [278, 279] and is initiated by ligand-receptor interaction between virus and host, which triggers an internalization signal. This signal recruits the scaffold protein clathrin to the plasma membrane and results in the electron-dense coat visible by electron microscopy (e.g. in Fig. 5.10B). The clathrin-coated vesicle is then pinched off from the plasma membrane by dynamin, uncoated during its transport to endosomes, fused to endosomes, and finally transformed into a lysosome [278]. The latter process involves acidification, which often causes the capsid to open and release the viral genome into the cytoplasm. Figure 5.10C shows a *C. roenbergensis* cell with two early CroV cores (arrow heads), multiple Mavirus particles attached to the cell surface, and a vesicle close to the plasma membrane containing a single Mavirus particle (arrow). Higher magnification of this area shows that the vesicle is coated with a 25-30 nm thick layer (Fig. 5.10D, arrow), as would be expected from CME [280]. An uncoated Mavirus-containing vesicle from a different cell is shown in Fig. 5.10E. These observations strongly suggest that the Mavirus virophage enters *C. roenbergensis* cells independently of CroV by a CME pathway. It cannot be excluded, however, that Mavirus can only attach to CroV-infected cells, e.g. if CroV infection triggered a biochemical response at the cell surface, which Mavirus could specifically detect.

Figure 5.10 Clathrin-mediated endocytosis of Mavirus.

(**A**) Mavirus virions attached to the cell surface. Scale bar = 100 nm. (**B**) Coated pit (arrow) formation at a site of attached Mavirus particles. Scale bar = 100 nm. (**C**) *C. roenbergensis* cell during early co-infection by CroV (arrow heads) and Mavirus (arrow). Scale bar = 1 µm. (**D**) Close-up view of the boxed area in (C) showing several extracellular Mavirus particles and a single virion inside a coated vesicle. The coat structure is marked by an arrow. Scale bar = 100 nm. (**E**) Intracellular uncoated vesicle containing a Mavirus particle. Scale bar = 100 nm. Symbols: CM, cytoplasma membrane; M, mitochondrion; MR, microtubular root; N, nucleus; RA, ribosome aggregates.



5.3.8 Mavirus replication within the CroV virion factory

No statement can be made about the length of the eclipse phase, as Mavirus virions were observed only in 24 h p.i. cells. This was probably due to a low MOI, and 1-2 rounds of Mavirus replication may have been necessary to generate sufficient progeny virions by 24 h p.i. for TEM visualization. Intracellular Mavirus particles were always located in the cytoplasm and never in the nucleus or other organelles. Furthermore, Mavirus particles were consistently found within and around the CroV VF, as shown in Fig. 5.11. The co-infected cell in Fig. 5.11A displays a large VF, several mature CroV particles, and Mavirus particles in the centre and along the periphery of the VF (arrowheads). In Fig. 5.11B, multiple virophage particles can be seen within the CroV VF. The individual particles had different densities, suggesting that Mavirus genome packaging may occur within the VF. It is therefore likely that replication and assembly of the virophage take place within the CroV VF, corroborating the virophage concept (see chapter 4 for further arguments). The presence of Mavirus in CroV VFs did not seem to directly affect CroV capsid assembly, as shown in Fig. 5.11C. However, some cells exhibited large VFs that contained Mavirus particles but no CroV virions (Figs. 5.11D, 5.13B), with the exception of occasionally observed objects that resembled abnormal CroV capsid structures (arrows in Fig. 5.11A,D; white arrowhead in Fig. 5.13B). In these cells, CroV replication may have been delayed or suppressed, a reasonable assumption given the decrease of CroV production observed in co-infected cultures (see section 4.3.1). Alternatively, the VF may not have yet started the production of CroV particles, which would imply that the Mavirus infection cycle is shorter than that of CroV. Mavirus has a much smaller genome (19 kbp) than CroV (730 kbp) and may therefore require less time for DNA replication. Similar results have been reported from the Sputnik virophage, which has a shorter replication cycle than its host virus Mimivirus [31].

Mavirus particles had a hexagonal cross-section which implies icosahedral morphology. With a diameter of ~60 nm, they were clearly distinguishable from the much larger CroV virions. Figure 5.12 visualizes the size difference between Mavirus and CroV. For comparison, the mitochondria in Fig. 5.12A and B are of equal size.

In co-infected cells, progeny Mavirus particles tended to form paracrystalline arrays at a peripheral cytoplasmic location (Fig. 5.13A,B). Many virions in these paracrystalline arrays had an electron-lucent appearance and may be incompletely packaged particles.

Figure 5.11 Mavirus particles co-localize with the CroV virion factory.

(**A**) Co-infected cell with a large VF that has produced several CroV virions. Mavirus particles are located within and around the CroV VF (arrowheads). The arrow marks a putative abnormal CroV capsid structure. Scale bar = 500 nm. Inset: Magnified view of the boxed area in (A) showing a CroV particle and a Mavirus (arrowhead) particle. Scale bar = 100 nm. (**B**) CroV VF with mature, partially assembled, and putative defective (arrow) CroV capsids at its periphery and multiple Mavirus particles in its centre (arrowheads). Scale bar = 200 nm. (**C**) Cytoplasmic boundary of a CroV VF showing that CroV capsid assembly (arrows) occurs in direct proximity to Mavirus particles (arrowheads). Scale bar = 200 nm. (**D**) Cell with a large VF but no visible CroV virions. Mavirus particles are marked by arrowheads and a putative defective CroV capsid structure is marked by an arrow. Scale bar = 500 nm. Symbols: M, mitochondrion; N, nucleus; VF, virion factory.

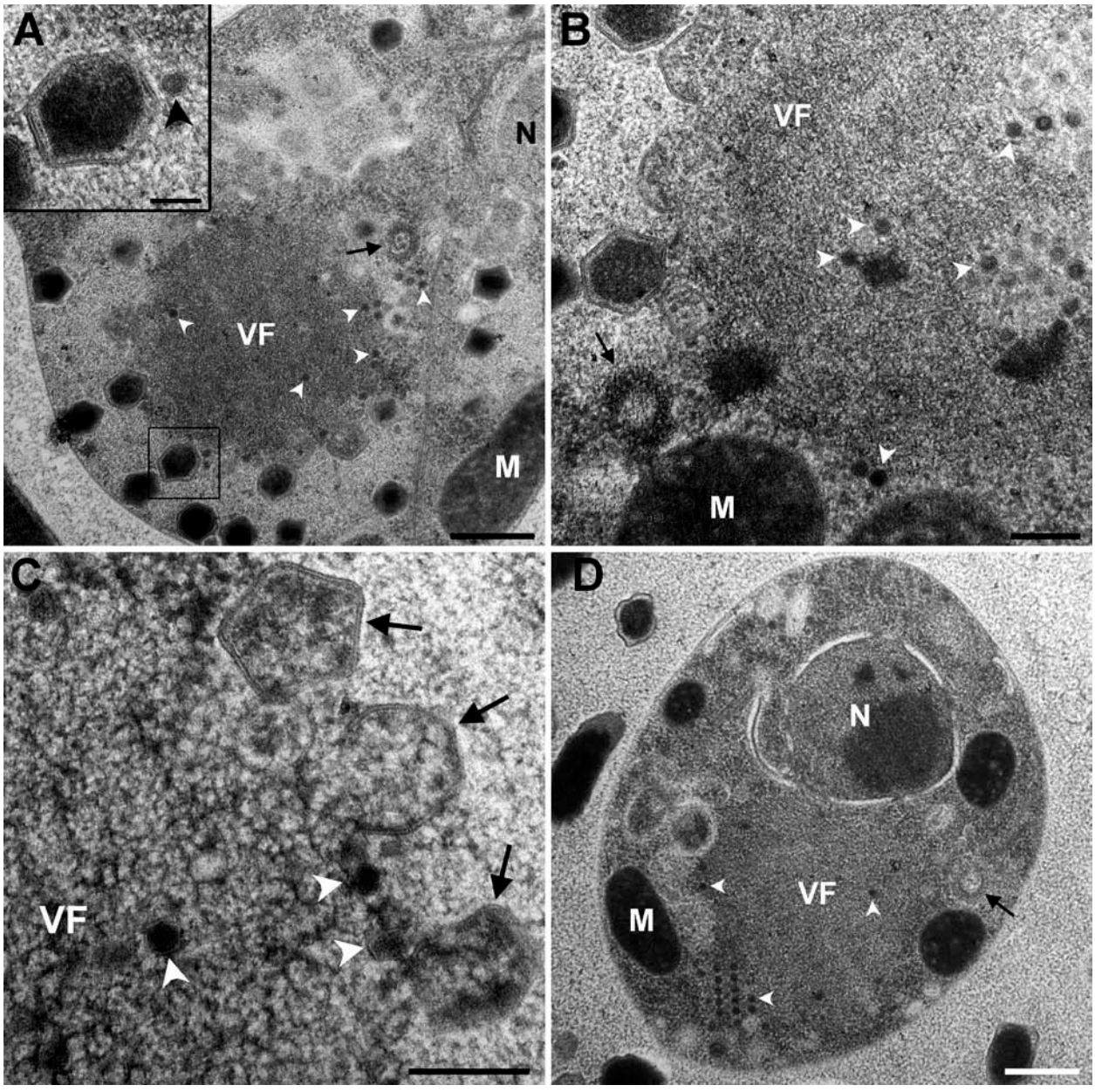
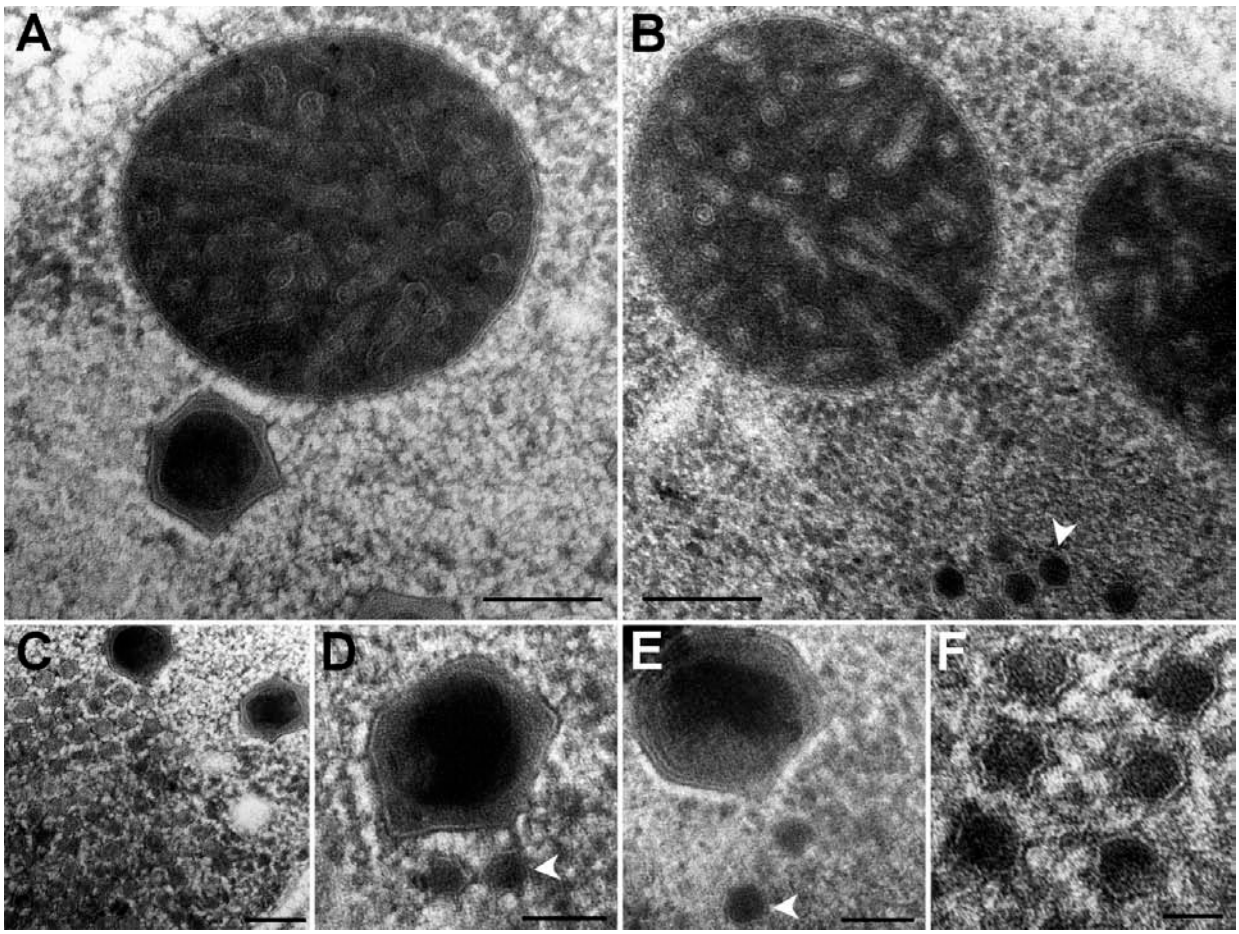


Figure 5.12 Intracellular virions of CroV and Mavirus.

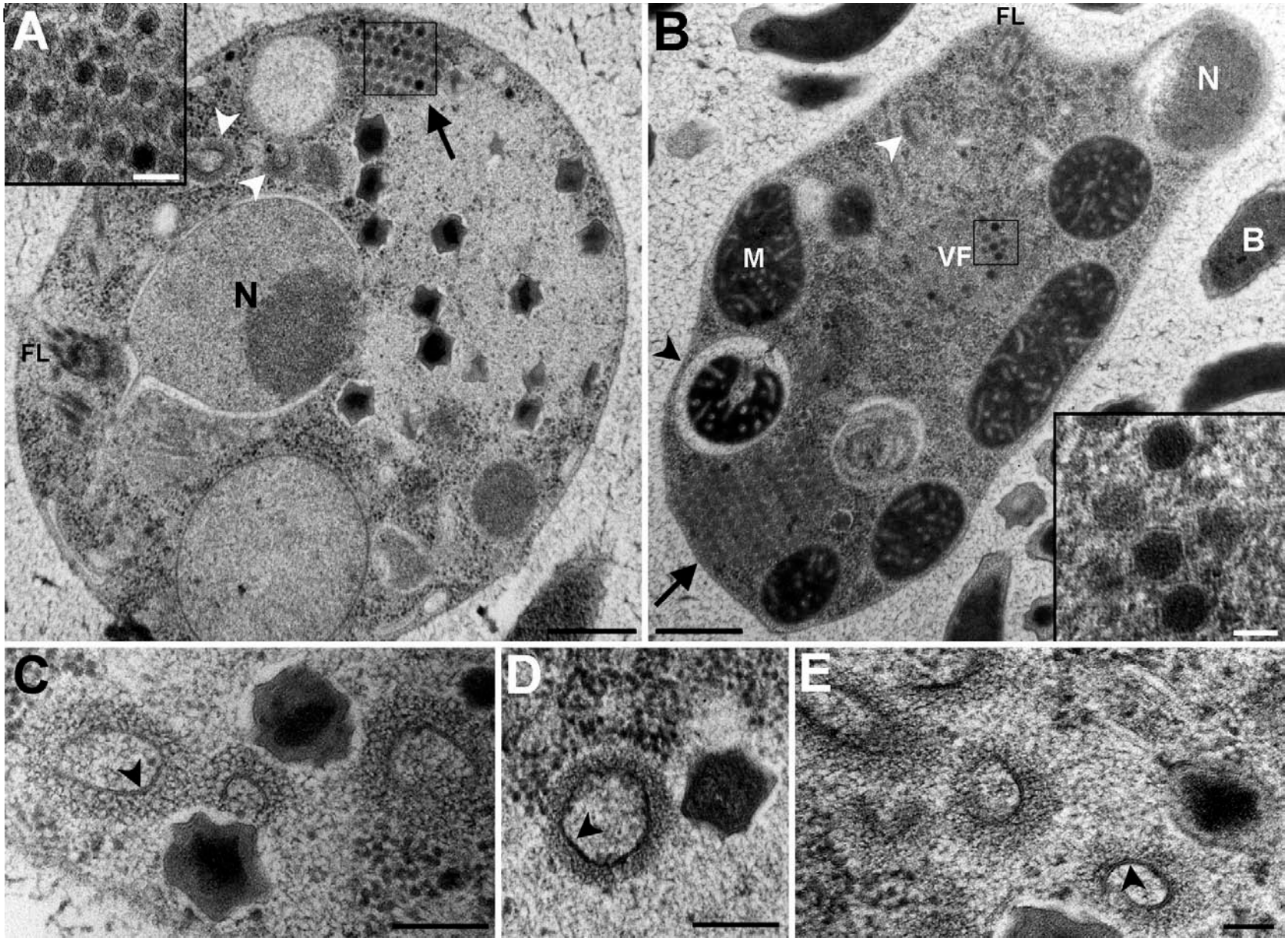
(**A**) CroV particle next to a mitochondrion. Scale bar = 200 nm. (**B**) Mavirus particles (arrowheads) next to mitochondria. Scale bar = 200 nm. (**C-E**) Mavirus particles (arrowheads) and CroV particles in direct proximity. Scale bars = 200 nm (C), 100 nm (D,E). (**F**) Close-up view of Mavirus virions. The hexagonal capsids contain electron-dense material. Scale bar = 50 nm.



Cells that were visibly infected with Mavirus frequently contained 100-400 nm large structures that consisted mostly of open circles (Fig. 5.13C,E), but closed circles were also observed (Fig. 5.13D). They were unilaterally covered with 35-70 nm long fibres, similar to the fibres that coat the CroV cores during early infection (Fig. 5.2). These fibres were anchored on a 7-9 nm thick electron-dense layer that resembled CroV capsid layers, with the exception that these structures lacked polygonal symmetry. Given that most of these structures were linear, it can be excluded that they are lipid membranes. Electron microscopy of *Acanthamoeba* cells co-infected with Mamavirus and Sputnik revealed defective Mamavirus capsids in the presence of Sputnik [31]. Some of these defective structures were elongated, open-ended capsid layers that were unilaterally covered with fibrils. It is therefore possible that the fibrous structures associated with Mavirus infection represent morphologically aberrant forms of CroV.

Alternatively, they may represent empty CroV capsids as they bear similarity to empty Mimivirus capsids in phagocytic vacuoles after release of the viral core into the cytoplasm [67]. They are also similar to virion structures of myxoma virus during stages of uncoating [281]. Unlike Mimivirus, however, CroV particles are not covered by external fibres and would have to develop these appendages upon contact with the host cytosol. Such a scenario is not impossible, given that even extracellular virion development has been documented in an archaeal virus [282]. As a further alternative, an outer fibre layer may have already been present in the common ancestor of CroV and Mimivirus, which was then reduced to a rudimentary stage in CroV. Because these structures were mostly observed in Mavirus-containing cells, it seems most likely, however, that they represent aberrant forms of CroV capsids caused by Mavirus co-infection.

Figure 5.13 Paracrystalline arrays of Mavirus and defective CroV capsid structures (next page). (**A**) Paracrystalline Mavirus array (arrow) and mature CroV particles in a co-infected cell. Putative defective CroV capsid structures are marked by arrowheads. The VF is not visible in this cross-section. Scale bar = 500 nm. Inset: Higher magnification of the paracrystalline array. Scale bar = 100 nm. (**B**) Mavirus-infected cell containing a relatively small VF, a large paracrystalline array of Mavirus particles (arrow), and a putative defective CroV capsid structure (white arrowhead). The nucleus is pushed to the cell periphery. The black arrowhead points to a phagocytosed mitochondrion, which was probably released from a lysed cell and ingested. Scale bar = 500 nm. Inset: Higher magnification of Mavirus particles in the VF. Scale bar = 50 nm. (**C-E**) Open and closed ring structures of putative abnormal CroV capsids (arrowheads) with attached fibres. Scale bars = 200 nm (C,D), 100 nm (E). Symbols: B, bacterium; FL, flagellar apparatus; M, mitochondrion; N, nucleus; VF, virion factory.


A
N
FL
B
FL
N
M
VF
B
C
D
E

5.4 Conclusions

CroV is an extraordinarily complex virus, encoding 544 predicted proteins and packaging at least 129 of them. Electron microscopy revealed that its intracellular mode of replication is similarly complex. Examination of ultra-thin sections of *C. roenbergensis* cells at different stages of infection provided the first ultrastructural description of the CroV replication cycle. CroV may enter the cell via endocytosis, as suggested by vesicles containing empty capsids. During the early stage of infection (starting immediately after inoculation), electron-dense CroV cores with a fibrous outer layer were found in the cytoplasm. As the dense core disappeared, the VF started to develop and grow, eventually surpassing the nucleus in size. Production of CroV virions started between 6 h p.i. and 12 h p.i. and entailed the fabrication of icosahedral capsids at the periphery of the VF, followed by a presumably two-stage process of genome packaging. Viral protein synthesis may be sustained by large ribosomal aggregates, a hitherto undescribed phenomenon. Mature virions detached from the VF and accumulated in electron-lucent areas around the VF until particles were released from the cell through lysis. Intact nuclei were observed throughout infection, suggesting that the host genome may not be degraded. These observations are reminiscent of ultrastructural studies in Mimivirus, which also described large cytoplasmic VFs with peripheral virion assembly via preformed capsids, and intact host nuclei during infection [67].

However, several questions remain unanswered, including the mode of cell entry by CroV. Uptake of the entire virion by endocytosis seems most likely, either via phagocytosis akin to Mimivirus or via receptor-mediated endocytosis. Furthermore, the origin of the internal virion membranes is unclear, as CroV VFs did not appear to be membrane-bound or associated with the host membrane system, a problem also described for Mimivirus [67]. The intracellular life cycle of CroV resembles that of Mimivirus and other nucleocytoplasmic large DNA viruses and suggests that CroV replicates exclusively in the host cytoplasm.

In addition to presenting the first detailed ultrastructural characterization of CroV infection, this study also provides the first visual description of the Mavirus virophage and is the most detailed report of the intracellular phase of a virophage thus far. Mavirus virions entered *C. roenbergensis* cells via clathrin-mediated endocytosis, were frequently found within the CroV VF but excluded from the nucleus, and formed paracrystalline arrays in the cytoplasm. Putative abnormal CroV capsid structures and a potential decrease in the number of mature CroV virions were observed in co-infected cells. These observations are in accordance with the recently proposed definition of a virophage, which states that virophages are replication-dependent on a co-infecting large DNA virus and are true parasites of the host virus, rather than the host cell.

This definition encompasses the idea that virophages replicate within the VF of the host virus, that they depend on gene products provided by the host virus, and that they decrease the fitness of their host virus as a result of parasitism.

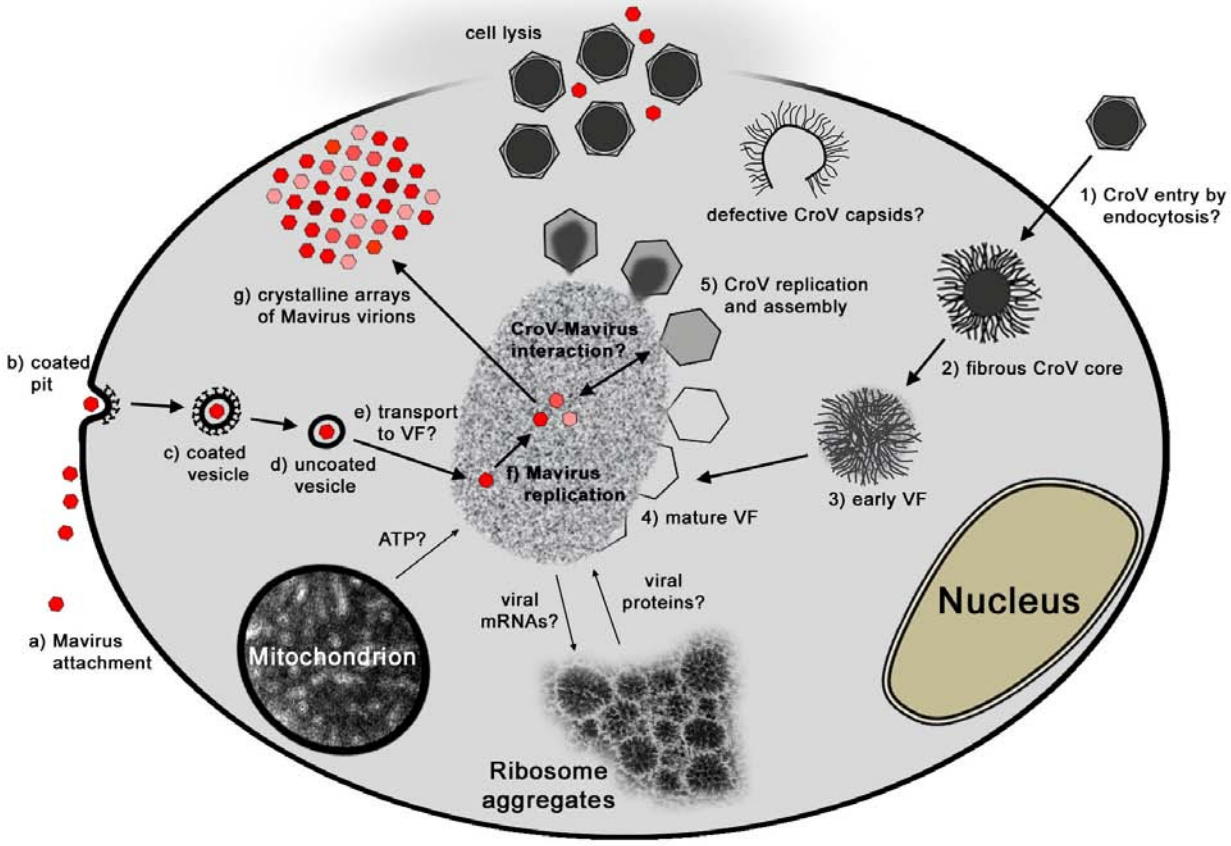
6 Concluding Chapter

In this dissertation, I presented a thorough characterization of a new giant virus, *Cafeteria roenbergensis* virus (CroV), and its virus parasite, Mavirus. By combining genomics, transcriptomics, proteomics, and thin-section transmission electron microscopy (TEM), a coherent picture about the biology of these viruses has started to emerge. Knowledge about genome sequence is the basis and key to further studies of any organism, especially one as small as a virus. Putative protein functions can be derived from a sequenced genome, which allow an assessment of the metabolic capabilities of the organism. Inferences about the evolutionary origin can be made from the analysis of regulatory regions and comparative phylogenetics. Furthermore, sequenced genomes allow the design of additional experiments, such as transcriptomics, proteomics, and characterization of individual gene products with biochemical methods. Such post-genomic studies are absolutely essential in order to obtain a comprehensive understanding of the biology of an organism. Although genome sequencing and bioinformatic methods are constantly being improved, caution should be exercised when protein functions are predicted on the sole basis of sequence homology. Accumulation of sequence data alone does not necessarily translate into a better understanding of gene function, as pointed out by Galperin and Koonin [283]. This should be kept in mind when interpreting the results of this dissertation. Despite the wealth of data generated from genome, transcriptome and proteome analyses, any statement about the biological function of a gene product is based on inference from sequence comparison and should be viewed as a prediction. Precautions have been taken to minimize erroneous predictions, e.g. by querying multiple databases, choosing conservative expectation value cutoffs, and identifying conserved protein domains via multiple sequence alignments. The strength of this work lies in the combination of data from different experiments and I hope that this information will be used as a starting point for further exploration of CroV, Mavirus, and related viruses. The following sections integrate and summarize the most important results from chapters 2-5 and discuss some important implications arising from the discovery of CroV and Mavirus.

6.1 The infection cycles of CroV and Mavirus

By combining genomics, thin-section electron microscopy, and temporal gene expression analysis, I was able to reconstruct key events of the CroV and Mavirus infection cycles. The intracellular steps of infection are summarized in Fig. 6.1 and outlined in the following sections.

Figure 6.1 The infection cycles of CroV and Mavirus in *C. roenbergensis*.



6.1.1 Before cell entry: the CroV particle

The viral particle is crucial for the dissemination of a virus, and analysis of the particle's ultrastructure and protein composition can give important clues about the biology and classification of a virus. The design of a viral capsid must meet high standards. In the extracellular milieu, the virion is exposed to harsh environmental conditions such as mechanical stress, UV irradiation, varying osmotic pressure, and reactive oxygen species. Only those viruses will succeed in evolution whose virions are able to protect the viral genome from these harmful conditions and deliver it unscathed to a suitable host cell. The CroV virion is a 220-280 nm large icosahedral particle (as measured by TEM), which may possess two unique vertices (see section 5.3.4). The capsid is built from the major capsid protein in combination with numerous other proteins, most of which have not yet been identified. Proteomic analyses revealed that at least 129 different virus-encoded proteins constitute the CroV virion. The capsid contains two lipid membranes. The outer membrane may be in close contact with the inner surface of the capsid layer, an interaction that could be mediated by several transmembrane virion proteins (Table 3.1). The inner membrane surrounds the core with a 730 kbp dsDNA

genome and an eclectic set of non-structural proteins. Among these proteins are at least six predicted DNA repair enzymes, involved in photoreactivation and base excision repair. CroV was isolated from the coastal waters of the Gulf of Mexico and the surface waters at this latitude are exposed to high doses of UV irradiation, hence it can be assumed that the viral DNA suffers considerable photodamage in its natural environment. It remains to be tested whether the packaged DNA repair enzymes are active inside the virion. The CroV particle is one of the most complex known virions, containing an elaborate set of structural and non-structural proteins, as well as a multi-layered internal structure.

6.1.2 Cell entry of CroV

During TEM analysis of infected cells, no case of virus attachment or penetration could be documented. However, vesicle-enclosed empty capsids in the cytoplasm of visibly infected cells suggested that CroV may enter its host by an endocytic pathway (see section 5.3.3). Whether CroV is internalized by receptor-mediated endocytosis or phagocytosis remains to be determined. During Mimivirus cell entry, the amoeba ingests the virus particle, the Mimivirus stargate portal opens inside the phagosome, and the outer virion membrane fuses with the phagosomal membrane, releasing the virion core into the cytoplasm (see section 1.2.2). CroV, however, lacks the fibrous capsid coating of Mimivirus, which presumably has the effect of mimicking the peptidoglycan layer of bacterial prey and increasing the effective particle diameter to trigger phagocytosis. Hence, receptor-mediated endocytosis seems to be more likely for CroV. Entry of CroV may be facilitated by virion-located oxidoreductases, protein disulfide isomerases, and proteases present in the virion, which could induce conformational changes upon contact with the host cell.

6.1.3 Early steps in CroV infection

The first visible sign of infection was the presence of electron-dense cytoplasmic vesicles that matched CroV cores in diameter and electron density (see section 5.3.2). CroV encodes an extensive array of transcription-related proteins and most of this machinery was found to be packaged inside the virus particle (see section 3.3.3). The CroV transcription apparatus includes eight RNA polymerase subunits, several transcription factors and helicases, DNA topoisomerase IB, the mRNA capping enzyme and a polymerase for mRNA polyadenylation. In this regard, CroV is similar to poxviruses and Mimivirus, where transcription of early genes takes place inside the virion cores before the genome is released into the host cytoplasm. It is therefore very likely that early transcription of CroV genes operates in a similar manner. According to this scenario, protein components required for transcription are packaged in the

virion core and the mRNAs of early genes are 5' capped and 3' polyadenylated within the confinement of the inner virion membrane. In TEM thin sections, the intracellular core-like vesicles during the early stages of infection were surrounded by a dense layer of fibrous material, which may consist of mRNA emerging from the viral core, targeted for translation in the cytoplasm.

Free ribosomes were present throughout the cytoplasm, but were mostly excluded from the virion factory (VF). Rosette-like structures that appeared to consist of aggregated ribosomes were frequently observed in CroV-infected cells, and also, albeit less frequently, in uninfected cells. The literature on *C. roenbergensis* contained no reference to such observations and it is not clear whether these structures can be induced by CroV.

Using microarray analysis, 62 CroV transcripts were detected immediately after infection (see section 2.3.10). Some of these transcripts may be packaged inside the virion. The early genes were mostly governed by the "AAAAATTGA" promoter signal, which may be recognized by packaged transcription factors. Thirty-nine of the genes expressed at 0 h p.i. were of unknown function, the remainder included diverse enzymatic functions such as ubiquitin ligases, heat shock proteins, SFII helicases, isoleucyl-tRNA synthetase, histone acetyl transferase, methyltransferases, a protein kinase, a protein phosphatase, and the large subunit of ribonucleotide reductase (RNR). It can be assumed that many of these early proteins are involved in virus-host interaction. Between 1 h p.i. and 3 h p.i., four RNA polymerase subunits, five transcription factors, and the mRNA capping enzyme were expressed. This implies that 1-2 hours after the virus enters the cell, the packaged transcription machinery is replaced by newly synthesized components. About 2-3 h p.i., DNA replication genes started to be expressed. This comprised the main processive enzyme DNA polymerase B, which was first expressed at 3 h p.i., as well as two DNA topoisomerases (types IA and IIA), primase, DNA ligase, three SFII helicases, clamp loader (PCNA), and a replication factor. At the same time, nucleotide metabolism genes were transcribed, including the small subunit of RNR, thymidine kinase, dUTPase, dNMP kinase, and the bifunctional dihydrofolate reductase/thymidylate synthase. These enzymes provide nucleotides for DNA replication. Other genes expressed at 1-3 h p.i. include a bacterial Lon protease, a Ser/Thr protein kinase, two ubiquitin ligases and a ubiquitin-specific protease, implying that the host metabolism may now be increasingly targeted by viral proteins. TEM thin sections of 3 h p.i. cells showed no sign of viral infection and VFs were difficult to identify due to their small size and lack of associated capsid structures.

6.1.4 CroV capsid assembly, genome packaging, and cell lysis

Between 3 h p.i. and 6 h p.i., CroV gene expression switches from early stage to late stage. A promoter signal associated with the late stage was found to contain a conserved "TCTA" core motif (see section 2.3.11). Structural components were first expressed at 6 h p.i., including the major capsid protein, capsid protein 2, the major core protein, and a phage tail collar domain protein. These genes are associated with the TCTA promoter, and their proteins, except for capsid protein 2, were highly abundant in the virion proteome (see section 3.3.2). The predicted DNA packaging ATPase of the FtsK-HerA family was first seen expressed at 6 h p.i., but was not found to be packaged. Disulfide bond formation in viral proteins is likely to be catalyzed by the suite of CroV-encoded redox proteins, analogous to the ER-independent redox pathway identified in vaccinia virus [237]. The first visual evidence of virion assembly was found at 12 h p.i.. Production of capsid structures and packaging of the viral genome thus started between 6 h p.i. and 12 h p.i., which was accompanied by growth of the VF (see section 5.3.3). Assembly of CroV capsid structures resembled that of Mimivirus and took place at the periphery of the VF. Single capsid faces appeared first, followed by multi-facet partial capsids and finally complete but empty capsids with pentagonal or hexagonal cross-sections. These empty capsid shells were still attached to the VF when electron-dense material, the viral genome, was injected into the pre-fabricated capsids. Mature virions eventually detached from the VF and accumulated in the cytoplasm until they were released by cell lysis.

Many of the events involved in capsid assembly and cell lysis remain unknown. For instance, nothing is known about the origin of the internal virion membranes, as the VF itself does not seem to be membrane-bound. Genome packaging probably involves the previously mentioned FtsK-HerA-like ATPase, but it is unclear whether the DNA is injected through one specific vertex, through one of the capsid faces, or through both as seen in Mimivirus [67]. Lysis occurred between 12 h p.i. and 18 h p.i., but the exact timing and the lysis-inducing factors are unknown.

In summary, CroV displays a highly complex and unusually autonomous infection cycle, which seems to occur entirely in the cytoplasm and resembles those of Mimivirus and poxviruses.

6.1.5 The infection cycle of the Mavirus virophage

The Mavirus virophage has an icosahedral capsid with a diameter of ~60 nm, as measured from TEM thin sections (Fig. 5.12). The existence of an inner membrane is suggested by a Mavirus homologue of the P9/A32-like, FtsK-HerA family of DNA pumping ATPases, which are

specifically found in dsDNA viruses with an internal lipid layer. The protein composition of the virion remains unknown, with the exception of the major capsid protein (MV18), which was identified by LC-MS/MS analysis during the proteome experiment (see section 3.3.9).

TEM thin sections of infected cells demonstrated that the virophage enters *C. roenbergensis* cells independently of CroV by clathrin-mediated endocytosis (see section 5.3.7). Mavirus virions were frequently seen attached to the extracellular side of the plasma membrane and enclosed within coated and uncoated intracellular vesicles (Fig. 5.10). Nothing is known about the eclipse phase, i.e. the events that happen after endocytosis and before new virophage particles appear in the CroV VF. Using quantitative real-time PCR (qPCR), no Mavirus genome replication was observed in cultures inoculated with a virophage preparation that had been filtered through 0.1 µm syringe filters to remove CroV particles (see section 4.3.1). In contrast, Mavirus did replicate in cells co-infected with CroV. Mavirus replication is therefore dependent on co-infection with CroV. Mavirus encodes a predicted protein-primed DNA polymerase B, and a superfamily 3 helicase was found to be homologous to the principal replicative helicases of other dsDNA viruses. Mavirus could thus be self-sufficient for DNA replication and rather depend on CroV enzymes for gene expression. This would be in accordance with the highly conserved TCTA promoter signal that is located approximately 15 nt upstream of all 20 Mavirus CDSs and is identical to a predicted late promoter motif in CroV.

While it cannot be excluded that Mavirus has a nuclear phase in its infection cycle, there is currently no evidence indicating nuclear involvement, and it must be assumed that Mavirus replication and assembly take place in the cytoplasmic VF of CroV. This in turn implies that the only transcription enzymes available to Mavirus are those encoded and packaged by CroV. Preliminary data suggests that early CroV transcription starts inside the virion core, thus any early transcription factors packaged by CroV would not be accessible to a cytoplasmically located virophage genome, unless it were able to penetrate the CroV core. When CroV infection proceeds to the late stage of infection (around 6 h p.i.), the CroV core is transformed into a VF. This suggests that CroV-encoded late transcription factors are more abundant and more widely distributed than early transcription factors, which could explain why Mavirus evolved to incorporate the CroV late promoter motif in its 5' upstream regions instead of an early CroV promoter motif. Similar findings from the Mimivirus-Sputnik system lead to the conclusion that virophages in general tend to exploit the late transcription machinery of their host viruses [69].

Due to the mixed infection system and the presumably low virophage inoculum that was used for the TEM experiment, virophage particles were only found in 24 h p.i. samples, which was the latest sample time point. Mavirus virions were most often found within or around the CroV VF, a

further indication that replication and assembly of Mavirus may occur in the CroV VF. Several cells exhibited large crystalline arrays of virophage particles outside the CroV VF, which could represent a form of temporary storage for progeny virophages.

In summary, genome, qPCR, and electron microscopy analyses indicate that Mavirus is a virophage of CroV, and that Mavirus replication depends on CroV-encoded enzymes.

6.2 CroV genome evolution

6.2.1 The origin of CroV genes

Genomes of large DNA viruses vary greatly in size, from 100 kbp to 1,200 kbp. Comparative phylogenomics performed by Iyer and coworkers revealed a core set of ~50 genes that is shared by most NCLDVs and was likely present in the NCLDV ancestor [49, 50, 52]. The coding potential of extant NCLDVs is much larger than this core set and ranges from 150 to 981 proteins [28, 69]. There are two principal explanations for this discrepancy. On one hand, the NCLDV ancestor could have possessed a relatively small genome, which has been extensively expanded over the course of evolution. On the other hand, if the genome of the NCLDV ancestor was similar in size to those of extant giant viruses, then gene loss, gene replacement, and reductive evolution would be the main driving forces of NCLDV genome evolution. Although gene loss was probably an important factor in shaping at least some of the smaller NCLDV genomes, most researchers favour the former scenario of massive gene acquisition by giant viruses. To circumvent the problem that the majority of NCLDV genes are ORFans and have no homologues in cellular genomes, phylogenetic analyses are often restricted to the smaller proportion of NCLDV genes with homologues in eukaryotes and bacteria [84, 90]. This results in a distorted view of NCLDV evolution, because insights gained from these limited analyses cannot be extrapolated to the entire genome of a giant virus [284]. Alternatively, the existence of viral ORFans is sometimes explained by accelerated viral mutation rates or the poor database coverage of cellular and viral sequences as a result of insufficient gene pool sampling. In contrast to these views, it has been shown that the mutation rates of NCLDV genes are quite comparable to those in eukaryotes [165] and that the number of viral ORFans does not decrease with increasing sequence information about cellular genomes [285]. The existence of viral ORFans therefore clearly contradicts the traditional view of viruses as "gene pickpockets" and argues for a more complex scheme to explain the evolution of NCLDVs. CroV is no exception, encoding 277 CDSs that have no significant sequence similarity to proteins in GenBank (Fig. 2.8).

However, even the evolutionary origins of viral genes with well-characterized homologues in cellular genomes may be more complicated than they seem at first glance. According to the traditional view, these genes are the result of horizontal gene transfer (HGT) from a eukaryote or bacterium to the virus. In phylogenetic trees, however, CroV homologues of cellular genes rarely cluster within a particular group of potential eukaryotic host organisms and instead branch frequently between eukaryotic and bacterial groups (e.g. Figs. 2.9, 2.11, 2.18, 2.28). Although it cannot be excluded that methodological problems such as long-branch attraction are responsible for these observations, and although the lack of sequence information about the *C. roenbergensis* genome represents a significant drawback of this study, my data analysis does not support a scenario of extensive gene acquisition by CroV via HGT from eukaryotic genomes.

NCLDVs not only contain well-conserved versions of eukaryotic genes, but also encode a significant number of genes related to bacteria. Giant viruses that infect phagotrophic eukaryotes such as *Acanthamoeba* or endosymbionts of phagotrophs such as *Chlorella sp*. (an endosymbiont of *Paramecium bursaria*) were found to have the highest proportion of bacterial genes [214]. The most likely explanation for this phenomenon is that viruses which infect bacteriovores are frequently exposed to DNA from ingested bacteria and, in the case of strictly cytoplasmic viruses such as Mimivirus, bacterial and viral DNA can share the same intracellular compartment. This led to the notion of amoebae and other phagotrophs as melting pots of evolution, mixing viral, bacterial, and eukaryal genomes [39, 286]. Filée and Chandler [214, 287] noted that bacterial genes in Mimivirus and some phycodnaviruses occurred in clusters at the genome extremities and that these genes may have been transferred concomitantly. Although located in the middle of the CroV genome, the 38 kbp region of putative bacterial origin represents the most impressive example of large-scale HGT in a giant virus thus far. The exceptional status of this region was corroborated by multiple observations. First, none of the genes in the 38 kbp region are associated with the conserved early or late promoter motifs, which cover the rest of the genome rather uniformly (Fig. 2.5). Second, not a single protein encoded by the 38 kbp region was found to be packaged in the virion (Fig. 3.1). Third, genes in this region have no homologues in Mimivirus, unlike the rest of the genome. And finally, those CDSs in the 38 kbp region with homologues in cellular organisms are predicted carbohydrate metabolism enzymes with predominantly bacterial affiliation (Table 2.5). Although these observations are in favour of a relatively recent acquisition of the 38 kbp genomic segment from a bacterial source, the rest of CroV genes are presumably very ancient, as outlined in the following section.

In summary, the analysis of the CroV genome presented in this work did not result in a simple explanation that would account for the puzzling gene composition of CroV. Instead, some parts of the genome seem to have originated from eukaryotic and bacterial genes, and other parts arose by gene duplication; however, the majority of CroV genes are unique to this virus and their evolutionary origin is unknown. The CroV genome will likely become an integral part of the constantly growing NCLDV data set, and will help to refine our understanding about the origin and evolution of giant viruses.

6.2.2 Indicators for an ancestral origin of CroV

As a member of the NCLDV clade, whose ancestor probably dates back to the origin of the eukaryotic domain about two billion years ago [52], CroV is certainly no recently emerged pathogen. These questions can only be addressed through detailed comparative phylogenomics, which are beyond the scope of this work. However, certain pieces of data indicate that CroV may be closer to the NCLDV ancestor than most other viruses in this group.

There are two types of DNA ligases found among NCLDVs, the NAD-dependent type and the ATP-dependent type. Their distribution pattern is complementary, i.e. no virus has been found to encode both NAD ligases [50]. Interestingly, although the ATP-dependent form is more widespread, phylogenetic analysis suggests that the NCLDV ancestor had an NAD-dependent DNA ligase which it may have acquired from a bacteriophage [175]. During the course of evolution, the NAD-dependent form was replaced on multiple occasions with the ATP-dependent form [175]. CroV, like Mimivirus, entomopoxviruses and some iridoviruses, encodes an NAD-dependent DNA ligase, hence the ancestral version is most likely still present in these viruses.

Another example is the multi-subunit DNA-dependent RNA polymerase encoded by CroV, consisting of at least eight proteins (see section 2.3.4). Based on the presence in most NCLDVs of the largest two RNA polymerase subunits (Fig. 2.31), it is thought that the NCLDV ancestor already possessed a complex RNA polymerase and that gene loss in several lineages (e.g. EhV-86 is the only phycodnavirus to encode RNA polymerase subunits) is responsible for the patchy distribution of these genes in extant NCLDVs [50]. The seemingly complete set of RNA polymerase subunits in CroV thus indicates that this virus has retained much of the genetic complement that is thought to have been present in the NCLDV ancestor.

One of the surprising findings of CroV genome analysis was the presence of four intein elements in genes with well-characterized homologues (see section 2.3.8). Although inteins are found in all three domains of life, their distribution is sporadic and their evolutionary origins are

unknown [192]. Pietrokovski [203] proposed that inteins might be as old as the cells in which they are found, and that inteins originally had a beneficial role to the cell, the loss of which caused their slow extinction. Clearly, there is much that we do not understand about intein biology, but if their patchy distribution across eukaryotes, bacteria, and archaea, their high sequence divergence, the absence of inteins in genomes of multicellular organisms, and the ancient origin of the proteins they are found in, are testimony of an early evolutionary origin of inteins, then any organism containing several inteins must be ancient itself [203]. CroV encodes four inteins, more than any other eukaryotic virus, but only one of them (DNA polymerase B intein) is also found in Mimivirus, despite the fact that all four exteins (i.e. the host proteins in which inteins are inserted) have close homologues in Mimivirus. One possible explanation is that the CroV inteins colonized their host proteins relatively recently; a scenario that would suggest that horizontal spread of inteins is a relatively frequent phenomenon. However, this contradicts the common view of intein evolution, which predicts that colonization of empty allelic sites is very rare, especially in asexually reproducing organisms, and that inteins may be slowly going extinct [203]. The missing homing endonuclease domain in the CroV RNA polymerase II Rpb2 intein and the degenerate endonuclease domains in DNA polymerase and ribonucleotide reductase inteins (Table 2.4) also argue against a recent acquisition of this intein. Therefore, the finding of four intein elements in CroV may indicate an ancient origin of this virus genome.

6.3 Taxonomic classification

Neither CroV nor Mavirus have been assigned yet to an existing viral family. A proposal has been submitted to the International Committee on Taxonomy of Viruses (ICTV) to create the genus *Cafeteriavirus* in the family *Mimiviridae* for the new species *Cafeteria roenbergensis virus*. The justification for the creation of a new genus states that CroV "has many unique features and is clearly distinct from *Acanthamoeba polyphaga* mimivirus, the only current member of the *Mimiviridae*. Although both viruses have very large genomes (730 kbp and 1.181 kbp) and share the same early promoter motif as well as many genes that are exclusive to this family, there are substantial differences in coding potential (544 vs. 981 protein-coding genes), late promoter motifs, genome organization, particle size (280 vs. 500 nm), and host range (marine flagellate vs. freshwater amoeba). All these observations argue strongly in favor of establishing the new genus *Cafeteriavirus* within the family *Mimiviridae*." As more giant viruses are being discovered which fill in the missing branches of the NCLDV tree, CroV may eventually be classified into its own family.

No ICTV proposal has yet been submitted for Mavirus classification. Virophages represent a

hitherto unseen type of satellite virus and it is likely that Sputnik and Mavirus will represent their own genera in a family of virophages that is yet to be created.

6.4 Limitations and outlooks

Elucidation of the CroV genome, transcriptome, and proteome, as well as the discovery of the Mavirus virophage and ultrastructural documentation of the infection cycle of these viruses have significantly expanded the current knowledge about giant viruses and their virophages, and have uncovered a genetic link between virophages and transposable elements. However, despite the wealth of new information gained, many aspects could not be sufficiently addressed within the scope of this thesis and several important questions remain unanswered.

First, it was not possible to sequence and assemble the CroV genome in its entire length. Large regions (~90 kbp at the 5' end and ~20 kbp at the 3' end) of repetitive DNA at the genome termini, combined with severe problems in the construction of large-insert genomic shotgun libraries, resulted in the technical inability to sequence these regions at present. As DNA sequencing technology continues to advance rapidly, it may be possible to elucidate the CroV genome termini in the future. This would allow a precise analysis of the repeat elements in the CroV genome and yield clues to their function, e.g., a role in genome replication or telomere-like functions.

Second, the CroV microarray experiment was performed before completion of genome sequencing and assembly, which resulted in the exclusion of CDSs from the array that were discovered later. Furthermore, validation of the microarray results through analysis of the virion proteome revealed that at least 44 viral mRNA species had not been detected by the microarray, although the corresponding proteins were found to be part of the virion (see section 3.3.6). The microarray results should therefore be interpreted with caution and used only to infer general trends, rather than to extract specific expression data about single genes. It seems worthwhile to conduct a more thorough transcriptomic analysis using 454 technology mRNA deep sequencing, because this technique provides information about transcriptional termination signals and has the potential to reveal new coding regions, as it is independent of a preselected set of hybridization probes [69].

Third, the Mavirus virophage was discovered during a late stage of the project and consequently, only a limited amount of time could be dedicated to its study. Most importantly, it was not possible to separate CroV quantitatively from Mavirus. A Mavirus-free CroV preparation is an imperative prerequisite for any future studies of the CroV-Mavirus system, since CroV

particle yield suffers significantly in the presence of Mavirus. Neither Optiprep-density gradient centrifugation nor preliminary experiments to selectively heat-inactivate Mavirus were successful in obtaining CroV preparations that were completely free of Mavirus particles. However, the separation was sufficient to demonstrate that CroV particle yields differ in the presence and absence of Mavirus. A nearly CroV-free Mavirus preparation was achieved by tangential flow filtration, followed by filtration through 0.1 µm pore size PVDF syringe filters (see section 4.2.1). In order to corroborate the viral transposogenesis hypothesis, the genome of *C. roenbergensis* should be scrutinized for integrated forms of Mavirus. This may be done best by subjecting the host cell genome to 454 pyrosequencing and searching for the Mavirus genome signature among the resulting contigs. Any contig that contains the Mavirus genome, linked to additional DNA sequence, would be proof for the existence of a provirophage. Ultimately, a provirophage-containing strain of *C. roenbergensis* could be used to study reactivation of Mavirus, e.g. in response to CroV superinfection. Such experimental evidence would demonstrate that a long-term association of host cell and virophage is beneficial to the eukaryotic host.

This dissertation will hopefully inspire biochemical and structural studies of CroV and Mavirus proteins. The role of CroV-encoded translation components could be a rewarding subject, e.g. by testing the hypothesis that Ile-tRNA synthetase, tRNAs and translation initiation factors such as eIF-4E initiate translation of CroV mRNAs at non-AUG start codons (AUA, AUU). Other potential target proteins include DNA repair enzymes, topoisomerases, histone acetyl transferase, tRNA pseudouridine 5S synthase, enzymes encoded by the 38 kb fragment of putative bacterial origin, protein disulfide isomerases and thioredoxins, and ubiquitin-pathway enzymes. In addition, the CroV genome may encode features suitable for biotechnological applications, e.g. inteins could be used for controlled protein splicing reactions or conserved promoter elements for developing new expression systems.

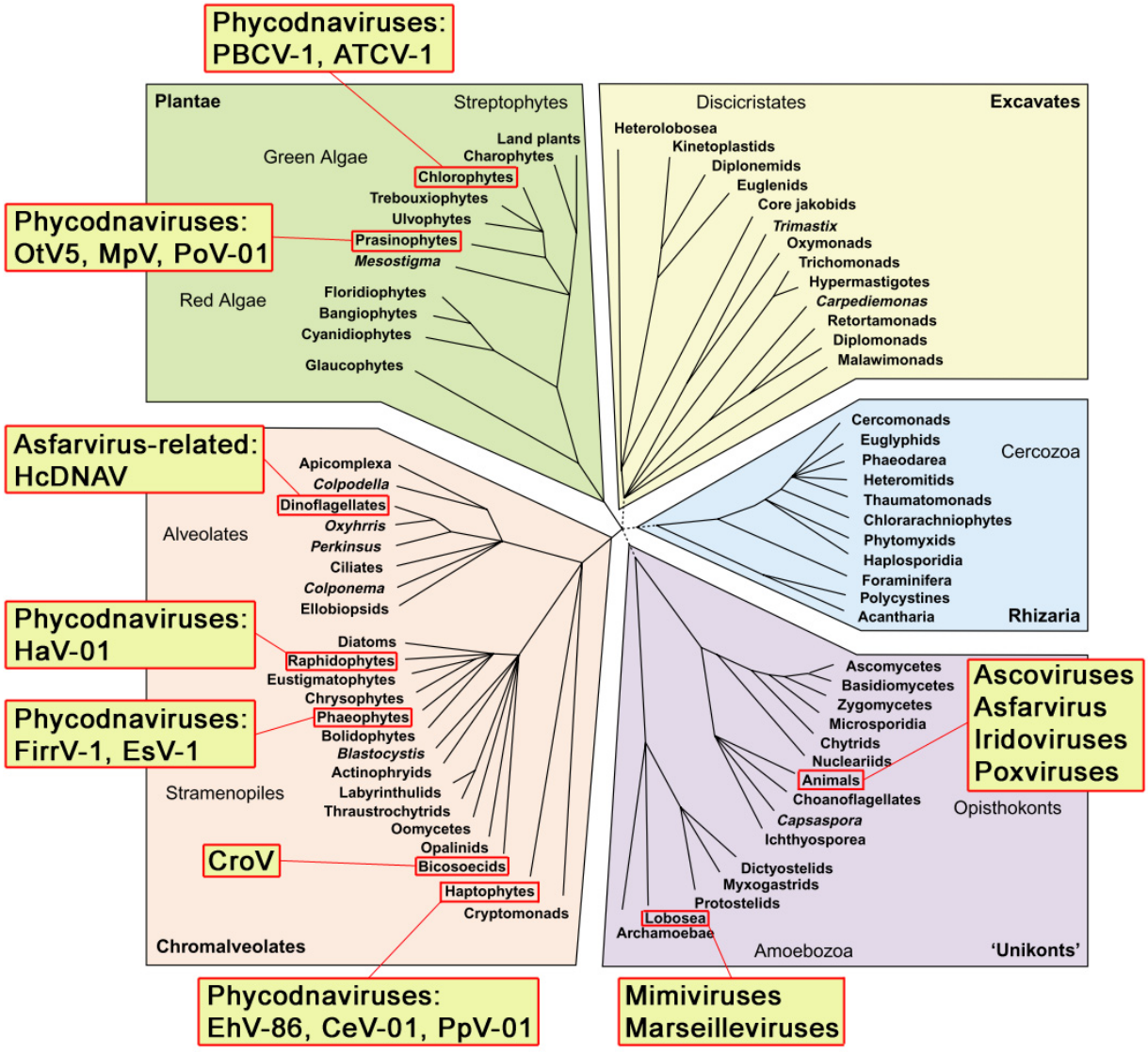
The ecology of CroV and Mavirus is a highly fascinating subject that could not be addressed within the scope of this dissertation. CroV is the first characterized virus of a ubiquitous marine zooplankton species and it can be inferred that this virus causes mortality in natural populations of *C. roenbergensis*, which implies an important role for CroV in nutrient cycling and for the community structure of microbial food webs. The ecological impact of the Mavirus virophage is difficult to gauge at this point. Virophages are parasites of giant viruses and benefit the cellular hosts that are infected by these giant viruses. Virophages are proposed to integrate into the genomes of their hosts (potentially both, host viruses and host cells), a process that most likely gave rise to an entire class of eukaryotic DNA transposons, the *Maverick/Polinton* elements. As

such, virophages are new players on the stage of microbial ecology and their exploration promises to reveal new ways in which viruses interact with each other and cellular organisms.

Finally, this work will inspire the search for new giant viruses. CroV represents the first virus infecting a bicosoecid and many other eukaryotic groups are still unexplored, including the viruses that may infect them. In Fig. 6.2, viruses of the NCLDV clade are paired with their host organisms on the tree of eukaryotes. Species of three eukaryotic supergroups (Plantae, Chromalveaolates, and Unikonts) are infected by NCLDVs and it is likely that under-sampling is the reason why no giant viruses are currently known to infect Excavates and Rhizaria. These lineages contain phagotrophic species of both ecological and medical importance, and given the widespread occurrence of NCLDVs, it is likely that a targeted search for giant viruses in new potential host organisms will be successful. As viruses co-evolve with their hosts and have a significant influence on cellular evolution (exemplified by the Mavirus-*Maverick/Polinton* relationship), the exploration of new viral hosts also holds great promise to discover new phenomena that will enrich our knowledge about early evolutionary events in the formation of cellular life [284].

Figure 6.2 Giant viruses and their hosts on the tree of eukaryotes.

Adapted from Keeling *et al.* [288], © Trends in Ecology & Evolution 2009 by permission from Elsevier.



6.5 Significance of the work

The discovery of CroV and the analysis of its 730 kbp genome represent a major advancement in the emerging field of giant virus research. As the first marine mimi-like virus and the largest known marine virus, CroV demonstrates that highly complex DNA viruses are found in oceanic environments, where they are likely to play ecologically important roles by infecting widespread marine zooplankton species such as *C. roenbergensis*. Analysis of environmental shotgun sequences had already predicted the existence of marine mimi-like viruses [97]; however, metagenomic approaches are no substitute for the isolation of specific virus-host systems, which allows for a much more detailed characterization.

With 544 predicted proteins, the CroV genome offers plenty of new targets for biochemical and bioinformatic studies. The finding of several translation- and metabolism-related genes demonstrates that these genes may be far more frequent among large DNA viruses than previously thought. Viral enzymes are often smaller than their cellular homologues and their biochemical characterization can help to identify "streamlined", minimal versions of a given catalytic activity [289, 290]. Phylogenetically, CroV represents an interesting case as it is clearly a member of the NCLDV clade with undeniable similarity to Mimivirus, although only a third of the CroV proteins appear to be related to Mimivirus proteins. On one hand, this provides a well-supported basis of common evolutionary origin for these two giant viruses. On the other hand, CroV and Mimivirus are different enough to examine the effects of virus-host gene transfer, the origin of bacterial genes in giant viruses, and the influence of lineage-specific gene expansion on viral genome evolution. Comparative genome analyses involving the CroV genome thus promise to significantly advance our knowledge about the evolutionary origins of giant viruses.

The discovery of the CroV-associated Mavirus virophage added another fascinating aspect to the biology of giant viruses. First, Mavirus highlights the unusual complexity of CroV by demonstrating that the intracellular virion factory of CroV contains sufficient resources to sustain the replication of another virus. Second, the CroV-Mavirus system shows that Mimivirus is not the only giant virus to be parasitized by a virophage and that Sputnik is not the only virus parasite of a giant virus. Sputnik was isolated from the freshwater environment of a cooling tower, whereas Mavirus most likely originated from the same coastal sea water sample that CroV had been isolated from. Hence, virophages are found in very different environments and a targeted search for them is likely to result in the identification of various new strains. Third, electron microscopy and co-infection studies of Mavirus and CroV confirmed several predictions of the virophage concept, such as dependency of virophage replication on a co-infecting host virus, co-localization of virophage particles with the virion factory of the host virus, and a profound negative effect of the virophage on host virus propagation [31]. Forth, despite causing similar phenotypes and exploiting similar strategies, such as replication within the virion factory of their host virus and a shared promoter signal with the host virus, Mavirus and Sputnik are genetically very different. Only four proteins among the two virophages show detectable sequence similarity and their gene order is not conserved. This implies that Sputnik and Mavirus have diverged from a common ancestor a very long time ago. Alternatively, the genetic differences may reflect the malleability of these genomes, if virophages are indeed novel vehicles of horizontal gene transfer among giant viruses [31]. The discovery of the Mavirus virophage had implications on an even grander biological scale. Only recently have we begun to appreciate the manifold and far-reaching impacts of viruses on cellular evolution. Viruses are a

source of novel genes [291] and during the course of evolution, viral genes have replaced cellular genes on multiple occasions, such as the bacterial RNA polymerase that was substituted with a bacteriophage version after endosymbiosis of an α-proteobacterium [292]. Our own genome bears witness of an evolutionary past that was heavily influenced by viruses, mainly retroid elements. Up to 45% of the human genome consists of transposable elements, and most of these are derived from retroviruses [128]. Genes of retroviral origin have been identified as key players in mammalian placentation [293], telomere maintenance [294], and chromosomal rearrangements [127]. Endogenous forms of non-retroviruses such as hepadnaviruses, bornaviruses, and filoviruses have recently been found in vertebrate genomes [295-297], and further viral links remain to be uncovered. In the context of these remarkable findings, the discovery of Mavirus revealed a completely new mechanism by which viruses have influenced the evolution of eukaryotic genomes. The viral transposogenesis hypothesis presented in this work states that the *Maverick/Polinton* DNA transposons originated from integration events of ancestral DNA virophages into eukaryotic genomes. This scenario is corroborated by the genetic similarities between Mavirus and *Maverick/Polinton* transposons and the mechanistic framework of virophage biology, which provides an evolutionary rationale for the association of a eukaryotic cell with a virophage. The ultimate origin of transposable elements is unknown and this work presents, for the first time, convincing data and a parsimonious evolutionary hypothesis for a viral origin of a class of DNA transposons.

Such remarkable findings would have been inconceivable at the beginning of this work. Despite being the most abundant biological entities on this planet and representing its largest genetic reservoir [7], marine viruses are still among the most understudied life forms and their exploration promises to uncover hitherto undescribed biological phenomena. The genetic complexity of CroV and its intricate and far-reaching interactions with the Mavirus virophage are prime examples for the incredible diversity of the viral world.

Bibliography

1. Lee RE (1971) Systemic viral material in cells of freshwater red alga *Sirodotia-Tenuissima* (Holden) Skuja*. J Cell Science* 8:623-631.

2. Bergh O, Borsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. *Nature* 340:467-468.

3. Proctor LM, Fuhrman JA (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* 343:60-62.

4. Suttle CA, Chan AM, Cottrell MT (1990) Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* 347:467-469.

5. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399:541-548.

6. Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* 64:69-114.

7. Suttle CA (2005) Viruses in the sea. *Nature* 437:356-361.

8. Fuhrman JA, Noble RT (1995) Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnol Oceanogr* 40:1236-1242.

9. Wilson WH et al. (2002) Isolation of viruses responsible for the demise of an *Emiliania huxleyi* bloom in the English Channel. *J Mar Biol Assoc UK* 82:369-377.

10. Wilson WH et al. (2005) Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science* 309:1090-1092.

11. Brussaard CPD (2004) Viral control of phytoplankton populations - a review. *J Eukaryot Microbiol* 51:125-138.

12. Nagasaki K, Tomaru Y, Shirai Y, Takao Y, Mizumoto H (2006) Dinoflagellate-infecting viruses. *J Mar Biol Assoc UK* 86:469-474.

13. Nagasaki K et al. (2005) Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl Environ Microbiol* 71:8888-8894.

14. Ogata H et al. (2009) Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virol J* 6:178-185.

15. Wilhelm SW, Suttle CA (1999) Viruses and nutrient cycles in the sea. *BioScience* 49:781-788.

16. Breitbart M (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99:14250-14255.

17. Suttle CA (2007) Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* 5:801-812.

18. Hambly E, Suttle CA (2005) The viriosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol* 8:1-7

19. Monier A et al. (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res* 19:1441-1449.

20. Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* 28:127-181.

21. Fauquet C, Mayo M, Maniloff J, Desselberger U, Ball L eds. (2005) Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses (Academic Press, San Diego). 2nd Ed.

22. Van Etten JL, Meints RH (1999) Giant viruses infecting algae. *Annu Rev Microbiol* 53:447-494.

23. Wilson WH, Van Etten JL, Allen MJ (2009) The Phycodnaviridae: the story of how tiny giants rule the world. *Curr Top Microbiol Immunol* 328:1-42.

24. Espagne E, Dupuy C, Huguet E (2004) Genome sequence of a polydnavirus: insights into symbiotic virus evolution. *Science* 306:286-289.

25. Kroemer JA, Webb BA (2004) Polydnavirus genes and genomes: emerging gene families and new insights into polydnavirus replication. *Ann Rev Entomol* 49:431-456.

26. Raoult D et al. (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* 306:1344-1350.

27. Claverie J-M et al. (2006) Mimivirus and the emerging concept of "giant" virus. *Virus Res* 117:133-144.

28. Schroeder DC et al. (2008) Genomic analysis of the smallest giant virus - *Feldmannia* sp. virus 158. *Virology* 384:223-232.

29. Van Etten JL, Lane LC, Dunigan DD (2009) DNA viruses: the really big ones (giruses). *Annu Rev Microbiol* 64:83-99.

30. La Scola B et al. (2010) Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology* 53:344-353.

31. La Scola B et al. (2008) The virophage as a unique parasite of the giant Mimivirus. *Nature* 455:100-104.

32. Fischer MG, Allen MJ, Wilson WH, Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 107:19508-19513.

33. Desjardins C et al. (2008) Comparative genomics of mutualistic viruses of Glyptapanteles parasitic wasps. *Genome Biol* 9:R183.

34. 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.

35. Hendrix RW (2009) Jumbo bacteriophages. *Curr Top Microbiol Immunol* 328:229-240.

36. Yan X, Chipman PR, Castberg T, Bratbak G, Baker TS (2005) The marine algal virus PpV01 has an icosahedral capsid with T=219 quasisymmetry. *J Virol* 79:9236-9243.

37. Baudoux A-C, Brussaard CPD (2005) Characterization of different viruses infecting the marine harmful algal bloom species *Phaeocystis globosa*. *Virology* 341:80-90.

38. 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.

39. 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.

40. Tulman ER et al. (2004) The genome of canarypox virus. *J Virol* 78:353-366.

41. Kapp M (1998) Viruses infecting marine brown algae. *Virus Genes* 16:111-117.

42. Müller DG, Wolf S, Parodi ER (1996) A virus infection in *Myriotrichia clavaeformis* (Dictyosiphonales, Phaeophyceae) from Argentinia. *Protoplasma* 193:58-62.

43. Delaroque N et al. (2001) The complete DNA sequence of the *Ectocarpus siliculosus* Virus EsV-1 genome. *Virology* 287:112-132.

44. Rohozinski J, Girton LE, Van Etten JL (1989) Chlorella viruses contain linear nonpermuted double-stranded DNA genomes with covalently closed hairpin ends. *Virology* 168:363-369.

45. Fitzgerald LA et al. (2007) Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect Chlorella Pbi. *Virology* 358:459-471.

46. Thomas JA et al. (2008) Characterization of *Pseudomonas chlororaphis* myovirus 201varphi2-1 via genomic sequencing, mass spectrometry, and electron microscopy. *Virology* 376:330-338.

47. Yang F et al. (2001) Complete genome sequence of the shrimp white spot bacilliform virus. *J Virol* 75:11811-11820.

48. Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. *Biol Direct* 1:29.

49. Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 75:11720-11734.

50. Iyer LM, Balaji S, Koonin EV, Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* 117:156-184.

51. 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.

52. Koonin EV, Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* 53:284-292.

53. Allen MJ, Schroeder DC, Holden MTG, Wilson WH (2006) Evolutionary history of the *Coccolithoviridae*. *Mol Biol* 23:86-92.

54. 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.

55. Schramm B, Locker JK (2005) Cytoplasmic organization of poxvirus DNA replication. *Traffic* 6:839-846.

56. Suhre K (2005) Gene and genome duplication in *Acanthamoeba polyphaga* mimivirus. *J Virol* 79:14095-14101.

57. Sicko-Goad L, Walker G (1979) Viroplasm and large virus-like particles in the dinoflagellate *Gymnodinium uberrimum*. *Protoplasma* 99:203-210.

58. Yamada T, Onimatsu H, Van Etten JL (2006) Chlorella viruses. *Adv Virus Res* 66:293-336.

59. Claverie J-M, Abergel C, Ogata H (2009) Mimivirus. *Curr Top Microbiol Immunol* 328:89-121.

60. La Scola B et al. (2003) A giant virus in amoebae. *Science* 299:2033.

61. Xiao C et al. (2005) Cryo-electron microscopy of the giant Mimivirus. *J Mol Biol* 353:493-496.

62. Xiao C et al. (2009) Structural studies of the giant Mimivirus. *PLoS Biol* 7:958-966.

63. Lien P, Pelkmans L, Capo C, Mege J-louis, Ghigo E (2008) Amoebal pathogen mimivirus infects macrophages through phagocytosis. *PLoS Pathog* 4:e1000087.

64. Ghigo E (2010) A dilemma for viruses and giant viruses: which endocytic pathway to use to enter cells? *Intervirology* 53:274-283.

65. Korn ED, Weisman RA (1967) Phagocytosis of latex beads by *Acanthamoeba*. II. Electron microscopic study of the initial events. *J Cell Biol* 34:219-227.

66. Zauberman N et al. (2008) Distinct DNA exit and packaging portals in the virus *Acanthamoeba polyphaga* mimivirus. *PLoS Biol* 6:e114.

67. Suzan-Monti M, La Scola B, Barrassi L, Espinosa L, Raoult D (2007) Ultrastructural characterization of the giant volcano-like virus factory of *Acanthamoeba polyphaga* mimivirus. *PLoS One* 2:e328.

68. Renesto P et al. (2006) Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. *J Virol* 80:11678-11685.

69. Legendre M et al. (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res* 20:664-674.

70. 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.

71. Byrne D et al. (2009) The polyadenylation site of Mimivirus transcripts obeys a stringent "hairpin rule". *Genome Res* 19:1233-1242.

72. Abergel C et al. (2005) Mimivirus TyrRS: preliminary structural and functional characterization of the first amino-acyl tRNA synthetase found in a virus. *Acta Crystallogr Sect F Struct Biol Cryst Commun* F61:212-215.

73. Abergel C, Giege R, Claverie J-M (2007) Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of Mimivirus TyrRS and MetRS. *J Virol* 81:12406-12417.

74. Jeudy S, Lartigue A, Claverie J-M, Abergel C (2009) Dissecting the unique nucleotide specificity of Mimivirus nucleoside diphosphate kinase. *J Virol* 83:7142-7150.

75. Jeudy S, Claverie J-M, Abergel C (2006) The nucleoside diphosphate kinase from Mimivirus: a peculiar affinity for deoxypyrimidine nucleotides. *J Bioenerg Biomem* 38:247-254.

76. Thai V et al. (2008) Structural, biochemical, and *in vivo* characterization of the first virally encoded cyclophilin from the Mimivirus*. J Mol Biol* 378:71-86.

77. Benarroch D, Qiu ZR, Schwer B, Shuman S (2009) Characterization of a Mimivirus RNA cap guanine-N2 methyltransferase. *RNA* 15:666-674.

78. Benarroch D, Smith P, Shuman S (2008) Characterization of a trifunctional Mimivirus mRNA capping enzyme and crystal structure of the RNA triphosphatase domain. *Structure* 16:501-512.

79. Benarroch D, Shuman S (2006) Characterization of Mimivirus NAD+-dependent DNA ligase. *Virology* 353:133-143.

80. Benarroch D, Claverie J-M, 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.

81. Bandaru V, Zhao X, Newton MR, Burrows CJ, Wallace SS (2007) Human endonuclease VIII-like (NEIL) proteins in the giant DNA Mimivirus. *DNA Repair* 6:1629-1641.

82. Monné M et al. (2007) The Mimivirus genome encodes a mitochondrial carrier that transports dATP and dTTP. *J Virol* 81:3181-3186.

83. Forterre P (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA* 103:3669-3674.

84. Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol Biol* 8:12-21.

85. Raoult D, Forterre P (2008) Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* 6:315-319.

86. Bandea CI (2009) The origin and evolution of viruses as molecular organisms. *Nature Precedings* hdl:10101/npre.2009.3886.1.

87. Koonin EV, Senkevich TG, Dolja VV (2009) Compelling reasons why viruses are relevant for the origin of cells. *Nat Rev Microbiol* 7:615.

88. Ludmir EB, Enquist LW (2009) Viral genomes are part of the phylogenetic tree of life. *Nat Rev Microbiol* 7:615.

89. López-García P, Moreira D (2009) Yet viruses cannot be included in the tree of life. *Nat Rev Microbiol* 7:615.

90. Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* 7:306-311.

91. Navas-Castillo J (2009) Six comments on the ten reasons for the demotion of viruses. *Nat Rev Microbiol* 7:615.

92. Raoult D (2009) There is no such thing as a tree of life (and of course viruses are out!). *Nat Rev Microbiol* 7:615.

93. Gowing MM (1993) Large virus-like particles from vacuoles of phaeodarian radiolarians and from other marine samples. *Mar Ecol Prog Ser* 101:33-43.

94. Gowing MM, Riggs BE, Garrison DL, Gibson AH, Jeffries MO (2002) Large viruses in Ross Sea late autumn pack ice habitats. *Mar Ecol Prog Ser* 241:1-11.

95. Venter JC et al. (2004) Sequencing of the Sargasso Sea. *Science* 304:66-74.

96. Rusch DB et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.

97. Ghedin E, Claverie J-M (2005) Mimivirus relatives in the Sargasso Sea. *Virol J* 6:1-6.

98. Monier A et al. (2008) Marine Mimivirus relatives are probably large algal viruses. *Virol J* 5:12.

99. Monier A, Claverie J-M, Ogata H (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* 9:R106.

100. Thomas V et al., Discovery and sequencing of a new nucleocytoplasmic large DNA virus. Poster Presentation, *Viruses of Microbes*, June 21-25 2010, Institut Pasteur, Paris, France.

101. 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.

102. Raoult D, Boyer M (2010) Amoebae as genitors and reservoirs of giant viruses. *Intervirology* 53:321-329.

103. Wommack KE, Sime-Ngando T, Winget DM, Jamindar S, Helton RR (2010) in Manual of Aquatic Viral Ecology, eds Wilhelm SW, Weinbauer MG, Suttle CA (American Society of Limnology and Oceanography), pp 110-117.

104. Sun S et al. (2010) Structural studies of the Sputnik virophage. *J Virol* 84:894-897.

105. Claverie J-M, Abergel C (2009) Mimivirus and its virophage. *Annu Rev Genet* 43:49-66.

106. Hassan F, Kamruzzaman M, Mekalanos JJ, Faruque SM (2010) Satellite phage TLCφ enables toxigenic conversion by CTX phage through dif site alteration. *Nature* 467:982-985.

107. Hewitt JA (1975) Miniphage-a class of satellite phage to M13. *J Gen Virol* 26:87-94.

108. Davis BM, Kimsey HH, Kane AV, Waldor MK (2002) A satellite phage-encoded antirepressor induces repressor aggregation and cholera toxin gene transfer. *EMBO J* 21:4240-9.

109. Christie GE, Calendar R (1990) Interactions between satellite bacteriophage P4 and its helpers. *Annu Rev Genet* 24:465-90.

110. Lindqvist BH, Dehò G, Calendar R (1993) Mechanisms of genome propagation and helper exploitation by satellite phage P4. *Microbiol Rev* 57:683-702.

111. Timpe JM, Verrill KC, Trempe JP (2006) Effects of adeno-associated virus on adenovirus replication and gene expression during coinfection. *J Virol* 80:7807-7815.

112. Samulski RJ (1993) Adeno-associated virus: integration at a specific chromosomal locus. *Curr Opin Genet Dev* 3:74-80.

113. Laughlin CA, Myers MW, Risin DL, Carter BJ (1979) Defective-interfering particles of the human parvovirus adeno-associated virus. *Virology* 94:162-174.

114. Gonçalves M a FV (2005) Adeno-associated virus: from defective virus to effective vector. *Virol J* 2:43.

115. Glansdorff N, Xu Y, Labedan B (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct* 3:29.

116. Mushegian A (2008) Gene content of LUCA, the last universal common ancestor. *Front Biosci* 13:4657-66.

117. Matthews R (1983) The origin of viruses from cells. *Int Rev Cytol Suppl* 15:245-280.

118. Bandea CI (1983) A new theory on the origin and the nature of viruses. *J Theor Biol* 105:591-602.

119. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. *Trends Microbiol* 8:504-508.

120. Prangishvili D, Stedman K, Zillig W (2001) Viruses of the extremely thermophilic archaeon Sulfolobus. *Trends Microbiol* 9:39-43.

121. Altenburg E (1946) The "viroid" theory in relation to plasmagenes, viruses, cancer and plastids. *Am Nat* 80:559-567.

122. Krupovic M, Bamford DH (2008) Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol* 6:941-948.

123. Benson SD, Bamford JK, Bamford DH, Burnett RM (1999) Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. *Cell* 98:825-833.

124. Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? *Curr Opin Struct Biol* 15:655-663.

125. Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. *Nature* 392:37-41.

126. Williamson SJ et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS One* 3:e1456.

127. Parseval N de, Heidmann T (2005) Human endogenous retroviruses: from infectious elements to human genes. *Cytogenet Genome Res* 110:318-332.

128. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci USA* 101:14572-14579.

129. Forterre P (2010) Defining life: the virus viewpoint. *Orig Life Evol Biophys* 40:151-160.

130. Claverie J-M (2006) Viruses take center stage in cellular evolution. *Genome Biol* 7:110.

131. Forterre P (2010) Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *Comptes Rendus Acad Sci*, in press.

132. Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* 3:537-546.

133. Patterson DJ (1991) The biology of free-living heterotrophic flagellates (Oxford University Press, Oxford [UK], New York).

134. Bettarel Y, Bouvy M, Arfi R, Amblard C (2005) Low consumption of virus-sized particles by heterotrophic nanoflagellates in two lakes of the French Massif Central. *Aquat Microb Ecol* 39:205-209.

135. 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.

136. Fenchel T, Patterson DJ (1988) *Cafeteria roenbergensis* nov. gen., nov. sp., a heterotrophic microflagellate from marine plankton. *Mar Microb Food Webs* 3:9.

137. 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.

138. 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.

139. Monger B, Landry M (1990) Direct-interception feeding by marine zooflagellates: the importance of surface and hydrodynamic forces. *Mar Ecol Prog Ser* 65:123-140.

140. Boenigk J, Arndt H (2002) Bacterivory by heterotrophic flagellates: community structure and feeding strategies. *Antonie Van Leeuwenhoek* 81:465-480.

141. Cavalier-Smith T (2006) Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). *J Mol Evol*:388-420.

142. St. John TM (2003) Characterization of a large DNA virus (BV-PW1) infecting the heterotrophic marine nanoflagellate *Cafeteria* sp. MSc Thesis, University of British Columbia.

143. Massana R, Campo J del, Dinter C, Sommaruga R (2007) Crash of a population of the marine heterotrophic flagellate *Cafeteria roenbergensis* by viral infection. *Environ Microbiol* 9:2660-2669.

144. Suttle CA, Fuhrman JA (2010) in Manual of Aquatic Viral Ecology, eds Wilhelm S, Weinbauer M, Suttle C (American Society of Limnology and Oceanography), pp 145-153.

145. Rutherford K et al. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16:944-945.

146. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276-277.

147. Altschul SF et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

148. Tatusov RL et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29:22-28.

149. Bateman A et al. (2002) The Pfam protein families database. *Nucleic Acids Res* 30:276-280.

150. Hunter S et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37:D211-D215.

151. 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.

152. 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.

153. 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.

154. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14:1188-1190.

155. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.

156. Dereeper A et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36:W465-W469.

157. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online - a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 33:W557-W559.

158. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.

159. Chou H (2010) Shared probe design and existing microarray reanalysis using PICKY. *BMC Bioinformatics* 11:196.

160. Allen MJ, Martinez-Martinez J, Schroeder DC, Somerfield PJ, Wilson WH (2007) Use of microarrays to assess viral diversity: from genotype to phenotype. *Env Microbiol* 9:971-982.

161. 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.

162. López-Lastra M et al. (2010) Translation initiation of viral mRNAs. *Rev Med Virol* 20:177-195.

163. Sonenberg N, Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136:731-745.

164. Cheung Y-nei et al. (2007) Dissociation of eIF1 from the 40S ribosomal subunit is a key step in start codon selection in vivo. *Genes Dev*:1217-1230.

165. Ogata H, Claverie J-M (2007) Unique genes in giant viruses: regular substitution pattern and anomalously short size. *Genome Res* 17:1353-1361.

166. Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* 17:1486-1495.

167. 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.

168. Oers MM van, Lampen MH, Bajek MI, Vlak JM, Eker APM (2008) Active DNA photolyase encoded by a baculovirus from the insect *Chrysodeixis chalcites*. *DNA Repair* 7:1309-1318.

169. 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.

170. Claverie J-M et al. (2009) Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* 101:172-180.

171. Pont-Kingdon GA et al. (1995) A coral mitochondrial mutS gene. *Nature* 375:109-111.

172. Lucas-Lledó JI, Lynch M (2009) Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Mol Biol Evol* 26:1143-1153.

173. Forterre P, Gribaldo S, Gadelle D, Serre MC (2007) Origin and evolution of DNA topoisomerases. *Biochimie* 89:427-446.

174. Chen F, Suttle CA (1996) Evolutionary relationships among large double-stranded DNA viruses that infect microalgae and other organisms as inferred from DNA polymerase genes. *Virology* 219:170-178.

175. Yutin N, Koonin EV (2009) Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. *Biol Direct* 4:51-64.

176. Van Etten JL, Lane LC, Meints RH (1991) Viruses and viruslike particles of eukaryotic algae. *Microbiol Rev* 55:586-620.

177. Komander D (2009) The emerging complexity of protein ubiquitination. *Biochem Soc Trans* 37:937-953.

178. Teale A et al. (2009) Orthopoxviruses require a functional ubiquitin-proteasome system for productive replication. *J Virol* 83:2099-2108.

179. Germot a, Philippe H (1999) Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. *J Eukaryot Microbiol* 46:116-24.

180. Peremyslov VV, Hagiwara Y, Dolja VV (1999) HSP70 homolog functions in cell-to-cell movement of a plant virus. *Proc Natl Acad Sci USA* 96:14771-14776.

181. Dolja VV, Kreuze JF, Valkonen JPT (2006) Comparative and functional genomics of closteroviruses. *Virus Res* 117:38-51.

182. Beere HM et al. (2000) Heat-shock protein 70 inhibits apoptosis by preventing recruitment of procaspase-9 to the Apaf-1 apoptosome. *Nature Cell Biol* 2:469-475.

183. Satyanarayana T et al. (2000) Closterovirus encoded HSP70 homolog and p61 in addition to both coat proteins function in efficient virion assembly. *Virology* 278:253-265.

184. Dolja VV, Karasev AV, Koonin EV (1994) Molecular biology and evolution of closteroviruses: sophisticated build-up of large RNA genomes. *Annu Rev of Phytopathol* 32:261-285.

185. Kolaczkowski B, Thornton JW (2009) Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PloS One* 4:e7891.

186. Brinker A et al. (2002) Ligand discrimination by TPR domains. Relevance and selectivity of EEVD-recognition in Hsp70 x Hop x Hsp90 complexes. *J Biol Chem* 277:19265-19275.

187. Blatch GL, Lässle M (1999) The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *BioEssays* 21:932-939.

188. Van Melderen L, Aertsen A (2009) Regulation and quality control by Lon-dependent proteolysis. *Res Microbiol* 160:645-651.

189. Ngo JK, Davies KJA (2007) Importance of the Lon protease in mitochondrial maintenance and the significance of declining Lon in aging. *Ann NY Acad Sci* 1119:78-87.

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

191. Frand AR, Cuozzo JW, Kaiser CA (2000) Pathways for protein disulphide bond formation. *Trends Cell Biol* 10:203-210.

192. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E (2002) Inteins: structure, function, and evolution. *Annu Rev Microbiol* 56:263-287.

193. Perler FB (2002) InBase: the intein database. *Nucleic Acids Res* 30:383-384.

194. Liu XQ (2000) Protein-splicing intein: genetic mobility, origin, and evolution. *Annu Rev Genet* 34:61-76.

195. Perler FB, Olsen GJ, Adam E (1997) Compilation and analysis of intein sequences. *Nucleic Acids Res* 25:1087-1093.

196. Ogata H, Raoult D, Claverie J-M (2005) A new example of viral intein in Mimivirus. *Virol J* 2:8.

197. 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.

198. 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.

199. Culley AI, Asuncion BF, Steward GF (2009) Detection of inteins among diverse DNA polymerase genes of uncultivated members of the *Phycodnaviridae*. *ISME J* 3:409-418.

200. Goodwin TJD, Butler MI, Poulter RTM (2006) Multiple, non-allelic, intein-coding sequences in eukaryotic RNA polymerase genes. *BMC Biol* 4:38-54.

201. Elleuche S, Pöggeler S (2010) Inteins, valuable genetic elements in molecular biology and biotechnology. *Appl Microbiol Biotechnol* 87:479-489.

202. Butler MI, Gray J, Goodwin TJD, Poulter RTM (2006) The distribution and evolutionary history of the PRP8 intein. *BMC Evol Biol* 6:42.

203. Pietrokovski S (2001) Intein spread and extinction in evolution. *Trends Genet* 17:465-472.

204. 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.

205. Cock JM et al. (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617-621.

206. Kobe B, Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* 11:725-732.

207. Catalano A, Luo W, Wang Y, O'Day DH (2010) Synthesis and biological activity of peptides equivalent to the IP22 repeat motif found in proteins from *Dictyostelium* and Mimivirus. *Peptides* 31:1799-1805.

208. Yanai-Balser GM et al. (2010) Microarray analysis of *Paramecium bursaria* chlorella virus 1 transcription. *J Virol* 84:532-542.

209. Grunwald-Beard L, Gamliel H, Parag G, Vedantham S, Zakay-Rones Z (1991) Killing of Burkitt-lymphoma-derived Daudi cells by ultraviolet-inactivated vaccinia virus. *J Cancer Res Clin Oncol* 117:561-567.

210. Tarahovsky YS, Ivanitsky GR, Khusainov AA (1994) Lysis of *Escherichia coli* cells induced by bacteriophage T4. *FEMS Microbiol Lett* 122:195-199.

211. Raetz CR (1990) Biochemistry of endotoxins. *Annu Rev Biochem* 59:129-170.

212. Brabetz W, Wolter FP, Brade H (2000) A cDNA encoding 3-deoxy-D-manno-oct-2-ulosonate-8-phosphate synthase of *Pisum sativum* L. (pea) functionally complements a kdsA mutant of the Gram-negative bacterium *Salmonella enterica*. *Planta* 212:136-143.

213. Filée J, Chandler M (2008) Convergent mechanisms of genome evolution of large and giant DNA viruses. *Res Microbiol* 159:325-31.

214. Filée J, Siguier P, Chandler M (2006) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet* 23:10-15.

215. Candiano G et al. (2004) Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. *Electrophoresis* 25:1327-1333.

216. 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.

217. Chan QWT, Howes CG, Foster LJ (2006) Quantitative comparison of caste differences in honeybee hemolymph. *Mol Cell Proteomics* 5:2252-2262.

218. Krogh A, Larsson B, Heijne G von, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567-580.

219. Braunagel SC, Russell WK, Rosas-Acosta G, Russell DH, Summers MD (2003) Determination of the protein composition of the occlusion-derived virus of *Autographa californica* nucleopolyhedrovirus. *Proc Natl Acad Sci USA* 100:9797-9802.

220. Loret S, Guay G, Lippé R (2008) Comprehensive characterization of extracellular herpes simplex virus type 1 virions. *J Virol* 82:8605-8618.

221. Huang C (2002) Proteomic analysis of shrimp white spot syndrome viral proteins and characterization of a novel envelope protein VP466. *Mol Cell Proteomics* 1:223-231.

222. Song W, Lin Q, Joshi SB, Lim TK, Hew C-L (2006) Proteomic studies of the Singapore grouper iridovirus. *Mol Cell Proteomics* 5:256-264.

223. Maxwell KL, Frappier L (2007) Viral proteomics. *Microbiol Mol Biol Rev* 71:398-411.

224. Allen MJ, Howard JA, Lilley KS, Wilson WH (2008) Proteomic analysis of the EhV-86 virion. *Proteome Sci* 6:1-6.

225. Resch W, Hixson KK, Moore RJ, Lipton MS, Moss B (2007) Protein composition of the vaccinia virus mature virion. *Virology* 358:233-247.

226. Wang R et al. (2010) Proteomics of the *Autographa californica* nucleopolyhedrovirus budded virions*. J Virol* 84:7233-7242.

227. Royo J, Gímez E, Hueros G (2000) CMP-KDO synthetase: a plant gene borrowed from gram-negative eubacteria. *Trends Genet* 16:432-433.

228. Iyer LM, Makarova KS, Koonin EV, Aravind L (2004) Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res* 32:5260-5279.

229. Mindich L, Bamford DH, McGraw T, Mackenzie G (1982) Assembly of bacteriophage PRD1: particle formation with wild-type and mutant viruses. *J Virol* 44:1021-1030.

230. Randall AZ, Baldi P, Villarreal LP (2004) Structural proteomics of the poxvirus family. *Artif Intell Med* 31:105-115.

231. Chung C-sheng, Chen C-hung, Ho M-yi, Huang C-yen (2006) Vaccinia virus proteome : identification of proteins in vaccinia virus intracellular mature virion particles. *J Virol* 80:2127-2140.

232. Yoder JD et al. (2006) Pox proteomics: mass spectrometry analysis and identification of vaccinia virion proteins. *Virol J* 3:10.

233. Mackinder LCM et al. (2009) A unicellular algal virus, *Emiliania huxleyi* virus 86, exploits an animal-like infection strategy. *J Gen Virol* 90:2306-2316.

234. Srinivasan V, Tripathy DN (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.

235. Schelhaas M et al. (2007) Simian Virus 40 depends on ER protein folding and quality control factors for entry into host cells. *Cell* 131:516-529.

236. Ryser HJ, Levy EM, Mandel R, DiSciullo GJ (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.

237. Senkevich TG, White CL, Koonin EV, Moss B (2002) Complete pathway for protein disulfide bond formation encoded by poxviruses. *Proc Natl Acad Sci USA* 99:6667-6672.

238. Fass D (2008) The Erv family of sulfhydryl oxidases. *Biochim Biophys Acta* 1783:557-566.

239. Gross E, Sevier CS, Vala A, Kaiser CA, Fass D (2002) A new FAD-binding fold and intersubunit disulfide shuttle in the thiol oxidase Erv2p. *Nat Struct Biol* 9:61-67.

240. Hakim M, Fass D (2010) Cytosolic disulfide bond formation in cells infected with large nucleocytoplasmic DNA viruses. *Antiox Redox Signal* 13:1261-1271.

241. Kates JR, McAuslan BR (1967) Poxvirus DNA-dependent RNA polymerase. *Proc Natl Acad Sci USA* 58:134-141.

242. Broyles SS (2003) Vaccinia virus transcription. *J Gen Virol* 84:2293-2303.

243. Van Vliet K et al. (2009) Poxvirus proteomics and virus-host protein interactions. *Microbiol Mol Biol Rev* 73:730-749.

244. Breitbart M et al. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99:14250-14255.

245. Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* 19:607-612.

246. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205-217.

247. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 41:95-98.

248. Gao X, Voytas DF (2005) A eukaryotic gene family related to retroelement integrases. *Trends Genet* 21:129-133.

249. Koonin EV, Zhou S, Lucchesi JC (1995) The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. *Nucleic Acids Res* 23:4229-4233.

250. Zhao Y et al. (1999) Crystal structure of an archaebacterial DNA polymerase. *Structure* 7:1189-1199.

251. Jurka J et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462-467.

252. Pritham EJ, Putliwala T, Feschotte C (2007) *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390:3-17.

253. Strömsten NJ, Bamford DH, Bamford JKH (2005) In vitro DNA packaging of PRD1: a common mechanism for internal-membrane viruses. *J Mol Biol* 348:617-629.

254. Desnues C, Raoult D (2010) Inside the lifestyle of the virophage. *Intervirology* 53:293-303.

255. Simossis V, Heringa J (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res* 33:W289-W294.

256. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.

257. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195-202.

258. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Ann Rev Genet* 41:331-368.

259. Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci USA* 103:4540-4545.

260. Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. *J Hered* 100:648-655.

261. Feschotte C, Pritham EJ (2004) Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet* 21:551-552.

262. Kulkosky J, Jones KS, Katz RA, Mack JP, Skalka AM (1992) Residues critical for retroviral integrative recombination in a region that is highly conserved among

retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Mol Cell Biol* 12:2331-2338.

263. Filée J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. *BMC Evol Biol* 13:1-13.

264. Iyer LM, Abhiman S, Aravind L (2008) A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. *Biol Direct* 3:39.

265. Pearson H (2008) “Virophage” suggests viruses are alive. *Nature* 454:677.

266. Johnson WE (2010) Endless forms most viral. *PLoS Genet* 6:e1001210.

267. O’Kelly CJ, Patterson DJ (1996) The flagellar apparatus of *Cafeteria roenbergensis* Fenchel & Patterson, 1988 (Bicosoecales=Bicosoeida). *Eur J Protistol* 32:216-226.

268. Novoa RR et al. (2005) Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biol Cell* 97:147-172.

269. Katsafanas GC, Moss B (2007) Colocalization of transcription and translation within cytoplasmic poxvirus factories coordinates viral expression and subjugates host functions. *Cell Host Microbe* 2:221-228.

270. Cherrier MV et al. (2009) An icosahedral algal virus has a complex unique vertex decorated by a spike. *Proc Natl Acad Sci USA* 106:11085-11089.

271. Castelló A et al. (2009) Regulation of host translational machinery by African swine fever virus. *PLoS Pathog* 5:e1000562.

272. Weinbauer MG, Suttle CA (1996) Potential significance of lysogeny to bacteriophage production and bacterial mortality in coastal waters of the Gulf of Mexico. *Appl Environ Microbiol* 62:4374-4380.

273. Etten V, L J, Burbank DE, Xia Y, Meints RH (1983) Growth cycle of a virus, PBCV-1, that infects Chlorella-like algae. *Virology* 126:117-125.

274. Derelle E et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS ONE* 3:e2250.

275. Nagasaki K, Tarutani K, Yamaguchi M (1999) Growth characteristics of *Heterosigma akashiwo* virus and its possible use as a microbiological agent for red tide control. Appl *Environ Microbiol* 65:898-902.

276. Doherty GJ, McMahon HT (2009) Mechanisms of endocytosis. *Annu Rev Biochem* 78:857-902.

277. Johannsdottir HK, Mancini R, Kartenbeck J, Amato L, Helenius A (2009) Host cell factors and functions involved in vesicular stomatitis virus entry. *J Virol* 83:440-453.

278. Mercer J, Schelhaas M, Helenius A (2010) Virus entry by endocytosis. *Annu Rev Biochem* 79:803-833.

279. Schelhaas M (2010) Come in and take your coat off - how host cells provide endocytosis for virus entry. *Cell Microbiol* 12:1378-1388.

280. Jin AJ, Prasad K, Smith PD, Lafer EM, Nossal R (2006) Measuring the elasticity of clathrin-coated vesicles via atomic force microscopy. *Biophys J* 90:3333-3344.

281. Duteyrat J-L, Gelfi J, Bertagnoli S (2006) Ultrastructural study of myxoma virus morphogenesis. *Arch Virol* 151:2161-80.

282. Häring M et al. (2005) Virology: independent virus development outside a host. *Nature* 436:1101-2.

283. Galperin MY, Koonin EV (2010) From complete genome sequence to “complete” understanding? *Trends Biotech* 28:398-406.

284. Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. *Intervirology* 53:362-378.

285. Forterre P, Prangishvili D (2009) The origin of viruses. *Res Microbiol* 160:466-472.

286. Moliner C, Fournier P-E, Raoult D (2010) Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol Rev* 34:281-294.

287. Filée J, Chandler M (2010) Gene exchange and the origin of giant viruses. *Intervirology* 53:354-361.

288. Keeling PJ et al. (2005) The tree of eukaryotes. *Trends Ecol Evol* 20:670-676.

289. Morehead TA et al. (2002) Ornithine decarboxylase encoded by Chlorella virus PBCV-1. *Virology* 301:165-175.

290. Gazzarrini S et al. (2009) Chlorella virus ATCV-1 encodes a functional potassium channel of 82 amino acids. *Biochem J* 420:295-303.

291. Daubin V, Ochman H (2004) Start-up entities in the origin of new genes. *Curr Opin Genet Dev* 14:616-619.

292. Filée J, Forterre P (2005) Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol* 13:510-513.

293. Blond JL et al. (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J Virol* 74:3321-3329.

294. Eickbush TH (1997) Telomerase and retrotransposons: which came first? *Science* 277:911-912.

295. Taylor DJ, Leach RW, Bruenn J (2010) Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol Biol* 10:193.

296. Horie M et al. (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463:84-87.

297. Gilbert C, Feschotte C (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol* 8:e1000495.

Appendices

Appendix A: 454 pyrosequencing and contig assembly

The CroV genome was sequenced using 454 high-throughput pyrosequencing. This method was chosen because it offered cost-effective and quick results and, more importantly, did not require the construction of shotgun clone libraries which posed a significant problem for the CroV genome. Datasets from two different 454 platforms were obtained: the first analysis in December 2005 used the early Genome Sequencer (GS) 20 platform; the second analysis in March 2009 was performed on the newer GS FLX system with Titanium chemistry. The main difference between these two datasets was the average read length, with the CroV GS FLX sequence reads being on average three times as long as the GS 20 sequences (370 bp vs. 119 bp). Although many of the GS 20 contigs had already been connected via Multiplex PCR and bioinformatic predictions before the GS FLX dataset became available, the GS FLX contigs were crucial in closing the last gaps and finalizing the assembly. In addition, the GS FLX analysis enabled the fortuitous discovery of the Mavirus virophage.

The final CroV genome assembly from GS 20 and GS FLX contigs is shown in Figure A.1. The longer read lengths of the GS FLX technology resulted on average in fewer and longer contigs, despite a few exceptions such as GS 20 contig 335, or the genome of an unknown marine phage (termed phage 307, Appendix E) which co-purified with CroV and was assembled into a single contig from the GS 20 data, whereas it was split into three contigs in the GS FLX dataset. In most cases, the contig ends differed between the two datasets, which facilitated assembly. However, larger regions of repetitive DNA could not be overcome by either dataset, resulting in identical contig ends. This was mostly the case for regions containing IP22 repeats, but also other, less common types of repeats and the tRNA gene cluster could not be assembled by the 454 software. Another problem caused by IP22 repeats were 'deletions' in the contigs, i.e. sequence repeats that were missing from the contigs, but were present in the CroV genome. These artificial deletions were caused by the assembly algorithm, which merged several repeat units into a single one. Nineteen such cases were uncovered (Table A.1) by comparison of the two datasets or by PCR product length anomalies and subsequent sequencing. It is possible that other occurrences may have gone unnoticed, especially if they were present at the same location in both 454 datasets. There were also several cases of misassembly, which resulted in the fusion of genomic regions that were in reality not adjacent to each other, and in the burial of 'true' contig breakpoints in the middle of a contig (see comments in Table A.1). Uncovering and pinpointing these assembly errors was often difficult and time-consuming, as the only reliable

hints to such occurrences were detailed partial contig alignments of the two 454 datasets and PCR analysis, i.e. PCR products whose existence was predicted by the assembly but could not be produced experimentally, or PCR products that were observed experimentally but were not predicted by the assembly.

Almost all 454 contigs were accounted for in the final assembly, except for a few short contigs that contained IP22 repeats, and two short, contaminating, bacterial contigs with low coverage (see Table A.1). This proves that the DNA samples that were used for 454 pyrosequencing were not contaminated by genomic DNA from the dense bacterial population present in the *C. roenbergensis* cultures and lysates. All of the large contigs containing unique, non-repetitive sequence were part of the final assembly, which led to the conclusion that the missing ends of the CroV genome are composed of repetitive DNA.

Figure A.1 Assembly map of the CroV genome showing the locations of 454 contigs.

Contigs are labelled by numbers, as allocated by the DNA sequencing facility. The scale bar corresponds to the CroV genome and is labelled in base pairs.

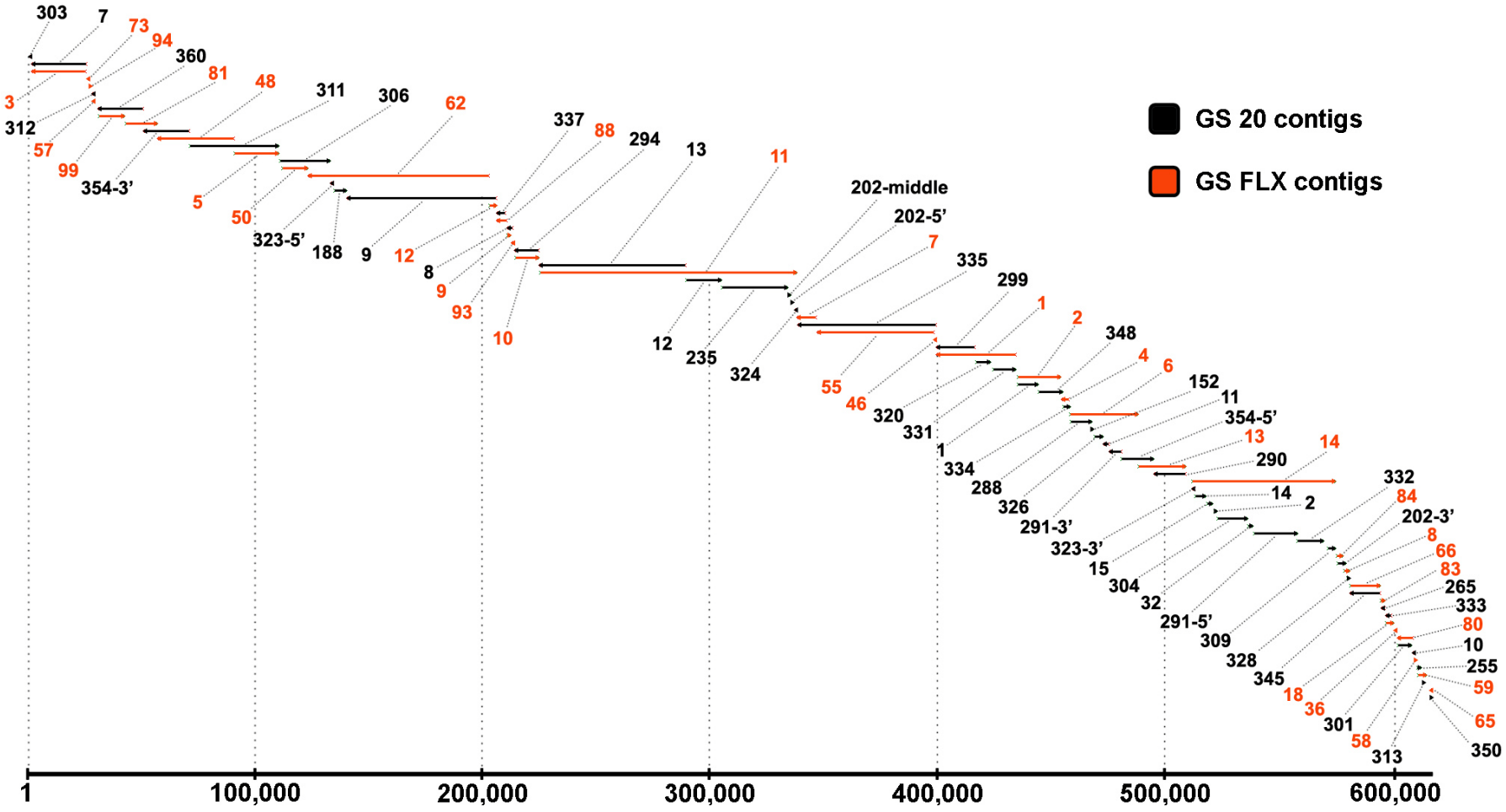


Table A.1 Contig statistics and location of contigs in the assembled CroV genome.

Deletions of more than 1 base pair are listed. Contigs listed in multiple parts were misassembled by the 454 assembly algorithm and had to be manually corrected.

454 contig	Length (bp)	Reads	Fold coverage	% A+T	CroV start	CroV end	Comments
GS20_001	9122	3004	39	76.23	435068	444190	
GS20_002	1376	511	44	79.29	521540	522915	
GS20_007	24032	6934	34	78.68	25911	1880	
GS20_008	2639	878	40	80.22	213518	210688	192 bp deletion
GS20_009	65787	20542	37	75.67	206101	140280	36 bp deletion
GS20_010	1100	396	43	76.27	609731	608632	
GS20_011	2523	710	33	77.80	475416	472894	
GS20_012	15799	4245	32	80.95	289272	305070	
GS20_013	64220	17888	33	77.32	289260	224981	60 bp deletion
GS20_014	5004	1641	39	78.06	513013	518016	
GS20_015	2599	801	37	78.68	518375	520973	
GS20_032	2468	700	34	78.85	536324	538791	
GS20_152	798	202	30	65.79	467885	468682	
GS20_188	5622	1690	36	76.65	134668	140289	
GS20_202	5017	1579	37	70.00	336401	336530	5' part
					334663	335584	Middle part
					575543	579507	3' part
GS20_235	29335	8762	36	77.02	305071	334406	
GS20_255	1877	1487	94	79.81	610727	612730	126 bp deletion
GS20_265	1537	493	38	77.36	596563	594961	66 bp deletion
GS20_288	9364	2856	36	76.22	458422	467785	

454 contig	Length (bp)	Reads	Fold coverage	% A+T	CroV start	CroV end	Comments
GS20_290	13906	4101	35	77.19	509122	495214	
GS20_291	24795	7136	34	76.98	538856	558187	5' part
					480892	475430	3' part
GS20_294	10586	3370	38	75.50	224823	214184	54 bp deletion
GS20_299	17124	5231	36	74.89	416634	399511	
GS20_301	6062	1960	38	80.53	602367	608428	
GS20_303	767	337	52	77.97	1489	657	132 bp deletion
GS20_304	13398	4002	36	77.98	522916	536313	
GS20_306	22245	6602	35	77.99	110822	133132	66 bp deletion
GS20_307	43363	10115	28	44.49	-	-	Phage307 genome
GS20_309	3772	885	28	50.64	571417	575188	
GS20_311	39396	11609	35	75.99	71191	110584	
GS20_312	1275	233	22	83.61	29777	28433	66 bp deletion
GS20_313	716	581	97	79.61	613195	613910	
GS20_320	6804	2021	35	78.42	416723	423526	
GS20_323	2710	982	43	78.56	134587	133093	5' part
					513064	511850	3' part
GS20_324	941	224	28	51.22	338121	337181	
GS20_326	4212	1333	38	76.02	468782	472993	
GS20_328	1436	440	36	75.14	579951	581386	
GS20_331	10105	3093	36	76.28	424381	434485	
GS20_332	11703	3635	37	75.67	558127	569829	
GS20_333	2779	793	34	76.07	599601	596823	
GS20_334	3199	971	36	74.99	455181	458379	

454 contig	Length (bp)	Reads	Fold coverage	% A+T	CroV start	CroV end	Comments
GS20_335	60772	19167	38	77.30	399342	338459	
GS20_337	3938	1293	39	74.05	210039	206102	
GS20_345	13255	3803	34	79.06	594664	581410	
GS20_348	10726	3391	38	76.06	444191	455180	264 bp deletion
GS20_350	849	725	102	84.81	616605	617453	
GS20_351	884	156	21	80.20	-	-	FNIP/IP22 repeats, aligns with GSFLX_82
GS20_354	34614	9639	33	78.35	480859	495257	5' part
					71267	50823	3' part; 228 bp deletion
GS20_360	19695	5483	33	77.93	50792	31098	
GSFLX_01	35006	3876	41	76.09	434534	399529	
GSFLX_02	19206	2150	41	76.11	435062	454269	
GSFLX_03	23950	2517	39	78.66	25868	1928	
GSFLX_04	3170	327	38	75.99	457802	454633	
GSFLX_05	19726	2133	40	75.78	90989	110716	
GSFLX_06	30108	3313	41	76.64	458082	488192	
GSFLX_07	8558	1022	44	78.59	346814	338256	
GSFLX_08	2675	249	34	77.50	578574	581248	
GSFLX_09	1554	233	55	80.18	210847	212400	
GSFLX_10	10512	1138	40	75.45	214320	224831	
GSFLX_11	112465	12390	41	77.15	225081	338081	534 bp deletion
GSFLX_12	3124	357	42	74.84	203081	206204	
GSFLX_13	20933	2452	43	77.97	488193	509127	
GSFLX_14	62916	6838	40	75.53	511850	575163	198 bp + 199 bp deletion
GSFLX_17	8160	130	6	44.44	-	-	Phage307 genome

454 contig	Length (bp)	Reads	Fold coverage	% A+T	CroV start	CroV end	Comments
GSFLX_18	3518	378	40	77.06	596892	600673	264 bp deletion
GSFLX_20	4814	82	6	44.85	-	-	Phage307 genome
GSFLX_24	534	2	1	43.26	-	-	60% aa identity to a heat shock protein GrpE from *Brucella pinnipedialis*, GrpE conserved domain of Hsp70, 59% aa identity to ribulose-phosphate 3-epimerase from *Ochrobactrum anthropi*
GSFLX_26	504	2	1	51.79	-	-	100% aa identity to Mg/Co transporter from *Idiomarina loihiensis*, cystathionine beta-synthase (CBS) conserved domain
GSFLX_36	537	64	44	76.54	601679	601011	132 bp deletion
GSFLX_46	1012	131	48	73.22	399373	398362	
GSFLX_47	29983	579	7	44.49	-	-	Phage307 genome
GSFLX_48	33705	3700	41	76.07	90844	57140	
GSFLX_50	11496	1323	42	79.41	111879	123376	
GSFLX_55	51359	5615	40	77.11	398350	346991	
GSFLX_57	885	131	55	85.08	29385	28501	
GSFLX_58	510	60	43	77.25	609829	610339	
GSFLX_59	3855	756	72	76.84	611219	615073	
GSFLX_60	18973	649	13	69.68	-	-	Mavirus virophage genome
GSFLX_62	79610	9014	42	75.90	203005	123393	
GSFLX_65	750	182	90	84.27	617342	616593	
GSFLX_66	13349	1414	39	79.07	581287	594635	
GSFLX_73	624	72	43	80.61	27084	26461	
GSFLX_79	1141	109	35	72.13	-	-	FNIP/IP22 repeats
GSFLX_80	6933	727	39	80.12	609216	602283	
GSFLX_81	14091	1601	42	78.06	42960	57171	120 bp deletion

454 contig	Length (bp)	Reads	Fold coverage	% A+T	CroV start	CroV end	Comments
GSFLX_82	863	92	39	81.69	-	-	FNIP/IP22 repeats, aligns with GS20_351
GSFLX_83	1681	157	34	76.98	594813	596559	65 bp deletion
GSFLX_84	2999	276	34	72.12	575330	578327	
GSFLX_87	534	55	38	79.21	-	-	FNIP/IP22 repeats
GSFLX_88	4349	489	41	74.68	210638	206197	93 bp deletion
GSFLX_93	1453	199	51	79.01	214154	212702	
GSFLX_94	811	73	33	79.53	27270	28080	
GSFLX_99	11579	1317	42	79.10	31150	42728	

Appendix B: Annotation of CroV CDSs

Table B.1 Functional annotations of CroV CDSs.

The column 'Microarray' lists the earliest time point (hours post infection) at which the respective transcript was detected in the microarray study (N/A = no microarray data available). The column 'Mimivirus' contains homologues of the respective CDSs in *Acanthamoeba polyphaga* mimivirus, as determined by BLASTP analysis. Mimivirus proteins in bold are confirmed virion components.

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov001	FNIP/IP22-like repeats	PF05725, FNIP, $2x10^{-18}$	N/A	early	-	6	L60, $1x10^{-12}$ and others	ADO67034
crov002	FNIP/IP22-like repeats	PF05725, FNIP, $4x10^{-19}$	2	early	-	6	L60, $3x10^{-81}$ and others	ADO67035
crov003	unknown	-	N/A	-	-	-	-	ADO67036
crov004	contains eIF-2B/eIF-5 domain	COG1601, GCD7, Translation initiation factor 2, beta subunit (eIF-2beta)/eIF-5 N-terminal domain [Translation, ribosomal structure and biogenesis] 0.001; PF01873, Domain found in IF2B/IF5, $3x10^{-4}$	N/A	-	-	-	-	ADO67037
crov005	unknown	-	N/A	early	-	-	-	ADO67038
crov006	unknown	-	-	-	-	837	-	ADO67039
crov007	unknown	-	12	early	-	-	-	ADO67040
crov008	unknown	-	2	early	-	-	-	ADO67041
crov009	unknown	-	0	-	-	-	-	ADO67042
crov010	tetratrico peptide repeats	COG0790 TPR repeat, SEL1 subfamily [General function prediction only] $5x10^{-9}$; PF08238, Sel1 repeat, $1x10^{-5}$	-	early	-	-	-	ADO67043
crov011	RNaseH	COG0258 Exo, 5'-3' exonuclease (including N-	6	early	-	-	-	ADO67044

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		terminal domain of PolI) [DNA replication, recombination, and repair] $4x10^{-9}$; PF02739, 5'-3' exonuclease, N-terminal resolvase-like domain, $9x10^{-10}$						
crov012	unknown	-	0	early	-	706	R588, $4x10^{-51}$	ADO67045
crov013	contains N-terminal oxidoreductase domain	COG0562, Pyridine nucleotide-disulphide oxidoreductase, 0.001; PF01266, FAD dependent oxidoreductase, $8x10^{-5}$	-	-	+	-	-	ADO67046
crov014	unknown	-	-	-	+	-	-	ADO67047
crov015	DNA N6-adenine methyltransferase	COG0286 HsdM, Type I restriction-modification system methyltransferase subunit [Defence mechanisms] $2x10^{-9}$; PF02384, N-6 DNA methylase, $4.1x10^{-6}$; PF01420, Type I restriction modification DNA specificity domain, $2x10^{-25}$	2	early	-	-	-	ADO67048
crov016	unknown	-	N/A	-	-	-	-	ADO67049
crov017	unknown	-	N/A	-	-	-	-	ADO67050
crov018	SFII helicase	COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination and repair] $2x10^{-8}$; PF04851, Type III restriction enzyme, res subunit, $4x10^{-4}$	6	early	-	-	-	ADO67051
crov019	unknown	-	-	-	+	-	-	ADO67052
crov020	unknown	-	-	-	-	-	-	ADO67053
crov021	unknown	-	1	early	-	-	-	ADO67054
crov022	tetratrico peptide repeats	COG0790, FOG: TPR repeat, SEL1 subfamily [General function prediction only] $6x10^{-5}$; PF08238, Sel1 repeat, $1x10^{-5}$	-	-	-	-	-	ADO67055
crov023	unknown	-	24	early	-	-	-	ADO67056

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov024	unknown	-	N/A	-	-	-	-	ADO67057
crov025	unknown	-	N/A	-	-	-	-	ADO67058
crov026	unknown	-	-	-	+	-	-	ADO67059
crov027	FNIP/IP22-like repeats	PF05725, FNIP, $4x10^{-19}$	N/A	early	-	6	L60, $4x10^{-36}$ and others	ADO67060
crov028	unknown	-	N/A	early	-	-	-	ADO67061
crov029	FNIP/IP22-like repeats	PF05725, FNIP, $1x10^{-15}$	N/A	early	-	6	L60, $2x10^{-15}$ and others	ADO67062
crov030	unknown	-	N/A	-	-	-	-	ADO67063
crov031	FNIP/IP22-like repeats	PF05725, FNIP, $1x10^{-18}$	2	early	-	6	L60, $9x10^{-60}$ and others	ADO67064
crov032	FNIP/IP22-like repeats	-	6	-	-	-	-	ADO67065
crov033	unknown	-	N/A	-	-	-	-	ADO67066
crov034	unknown	-	-	-	+	-	R871, $3x10^{-7}$	ADO67067
crov035	unknown	-	-	-	-	-	-	ADO67068
crov036	Lon protease	COG0466, Lon, ATP-dependent Lon protease, bacterial type [Posttranslational modification, protein turnover, chaperones] $4x10^{-99}$; PF05362, Lon protease (S16) C-terminal proteolytic domain, $1x10^{-31}$; PF00004, ATPase family associated with various cellular activities (AAA), $4x10^{-7}$	2	early	-	228	L251, $9x10^{-95}$	ADO67069
crov037	DNA endonuclease (XPG domain)	COG0258, Exo, 5'-3' exonuclease (including N-terminal domain of PolI) [DNA replication, recombination, and repair] $7x10^{-12}$; PF00752, XPG N-terminal domain, $4x10^{-26}$; PF00867,	-	-	-	1060	L386, $2x10^{-31}$	ADO67070

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		XPG I-region, $3x10^{-21}$						
crov038	5'-3' exoribonuclease	COG5049, XRN1, 5'-3' exonuclease [DNA replication, recombination, and repair / Cell division and chromosome partitioning / Translation] $2x10^{-28}$; PF03159, XRN 5'-3' exonuclease N-terminus, $5x10^{-14}$	-	-	-	1045	**R527/R528, $3x10^{-32}$**	ADO67071
crov039	unknown	-	-	late	+	-	-	ADO67072
crov040	unknown	-	N/A	-	-	-	-	ADO67073
crov041	unknown	-	6	early	-	-	-	ADO67074
crov042	unknown	-	N/A	early	-	-	-	ADO67075
crov043	ATP-dependent DNA helicase RecQ	COG0514, RecQ, Superfamily II DNA helicase [DNA replication, recombination, and repair] $2x10^{-87}$; PF00270 and, DEAD/DEAH box helicase, $1x10^{-32}$; PF00271, Helicase conserved C-terminal domain, $2x10^{-18}$	2	early	-	29	R290, $1x10^{-76}$ R458, $2x10^{-7}$	ADO67076
crov044	unknown	-	N/A	early	-	-	-	ADO67077
crov045	unknown	-	6	-	-	-	-	ADO67078
crov046	contains macro domain	-	N/A	-	-	-	-	ADO67079
crov047	unknown	-	6	late	+	-	-	ADO67080
crov048	exodeoxyribonuclease VII large subunit	COG1570, XseA, Exonuclease VII, large subunit [DNA replication, recombination, and repair] $2x10^{-39}$; PF02601, Exonuclease VII large subunit, $2x10^{-38}$	-	-	-	-	-	ADO67081
crov049	unknown	-	24	early	-	-	-	ADO67082
crov050	unknown	-	N/A	-	-	-	-	ADO67083
crov051	alkylated DNA repair	COG3145 AlkB, Alkylated DNA repair protein	-	early	-	-	R406, $1x10^{-8}$	ADO67084

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	protein	[DNA replication, recombination, and repair] $6x10^{-9}$; PF03171, 2OG-Fe(II) oxygenase superfamily, $2x10^{-9}$						
crov052	unknown	-	-	-	-	-	-	ADO67085
crov053	contains eIF-2B/eIF-5 domain	COG1601, GCD7, Translation initiation factor 2, beta subunit (eIF-2beta)/eIF-5 N-terminal domain [Translation, ribosomal structure and biogenesis] $2x10^{-11}$; PF01873, Domain found in IF2B/IF5, $3x10^{-5}$	6	early	-	-	-	ADO67086
crov054	unknown	-	6	-	-	-	-	ADO67087
crov055	unknown	-	-	-	-	-	-	ADO67088
crov056	DNA polymerase III, alpha subunit	COG2176, PolC, DNA polymerase III, alpha subunit (gram-positive type) [DNA replication, recombination, and repair] 0.005	6	-	-	-	-	ADO67089
crov057	unknown	-	12	-	-	-	-	ADO67090
crov058	ubiquitin-conjugating enzyme E2	COG5078 Ubiquitin-protein ligase [Posttranslational modification, protein turnover, chaperones] $5x10^{-31}$; PF00179, Ubiquitin-conjugating enzyme, $3x10^{-55}$	6	early	-	329	R795, $3x10^{-16}$ L460, $1x10^{-12}$ R521, $1x10^{-12}$ L709, $1x10^{-6}$	ADO67091
crov059	unknown	-	0	early	-	-	-	ADO67092
crov060	unknown	-	48	-	+	-	-	ADO67093
crov061	3'-5' exonuclease / tRNA(Ile) lysidine synthetase	COG0037, MesJ, Predicted ATPase of the PP-loop superfamily implicated in cell cycle control [Cell division and chromosome partitioning] $2x10^{-18}$; PF01171, PP-loop family, $5x10^{-24}$	2	early	-	1281	-	ADO67094
crov062	unknown	-	12	late	-	-	-	ADO67095
crov063	contains SUI1/eIF-1 domain	COG0023, SUI1, Translation initiation factor 1 (eIF-1/SUI1) and related proteins [Translation,	24	early	-	-	-	ADO67096

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		ribosomal structure and biogenesis] $5x10^{-9}$; PF01253, Translation initiation factor SUI1, $2x10^{-9}$						
crov064	unknown	PF00098, Zinc knuckle, $2x10^{-4}$	2	early	-	-	-	ADO67097
crov065	contains eIF-2B/eIF-5 domain	COG1601, GCD7, Translation initiation factor 2, beta subunit (eIF-2beta)/eIF-5 N-terminal domain [Translation, ribosomal structure and biogenesis] $8x10^{-4}$; PF01873, Domain found in IF2B/IF5, $8x10^{-11}$	72	early	-	-	-	ADO67098
crov066	FNIP/IP22-like repeats	PF05725, FNIP, $2x10^{-7}$	N/A	-	-	6	R636, $6x10^{-9}$ L60, $1x10^{-7}$	ADO67099
crov067	metalloendopeptidase	COG3590 PepO, Predicted metalloendopeptidase [Posttranslational modification, protein turnover, chaperones] $6x10^{-132}$; PF01431, Peptidase family M13, $6x10^{-69}$	6	-	+	663	R519, $3x10^{-26}$	ADO67100
crov068	FNIP/IP22-like repeats	PF05725, FNIP, $4x10^{-13}$	24	-	-	6	L60, $7x10^{-51}$ and others	ADO67101
crov069	deoxyuridine 5'-triphosphate nucleotidohydrolase	COG0756, Dut, dUTPase [Nucleotide transport and metabolism] $2x10^{-12}$; PF00692, dUTPase, $4x10^{-5}$	1	early	-	1068	-	ADO67102
crov070	unknown	-	N/A	-	-	-	-	ADO67103
crov071	bacterial tRNA pseudouridine 5S synthase	COG0130 TruB, Pseudouridine synthase [Translation, ribosomal structure and biogenesis] $4x10^{-7}$; PF01509, TruB family pseudouridylate synthase (N terminal domain), 0.033	1	early	-	-	-	ADO67104
crov072	SFII helicase	COG0553, HepA, Superfamily II DNA/RNA helicases, SNF2 family [Transcription / DNA replication, and repair] $2x10^{-12}$; PF00097; Zinc	0	early	-	330	R592, $3x10^{-15}$	ADO67105

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		finger, C3HC4 type (RING finger), $2x10^{-4}$; PF00176, SNF2 family N-terminal domain, $1x10^{-9}$						
crov073	unknown	-	-	early	-	234	-	ADO67106
crov074	DNA N6-adenine methyltransferase	COG0030 KsgA, Dimethyladenosine transferase (rRNA methylation [Translation, ribosomal structure and biogenesis] $2x10^{-4}$; PF00398, Ribosomal RNA adenine dimethylase, 0.022	24	late	-	-	-	ADO67107
crov075	asparagine synthetase B	COG0367 AsnB, Asparagine synthase (glutamine-hydrolyzing) [Amino acid transport and metabolism] $1x10^{-70}$; PF00733, Asparagine synthase, $2x10^{-82}$; PF00310, Glutamine amidotransferases class-II, $3x10^{-6}$	2	early	-	61	R475, $2x10^{-93}$	ADO67108
crov076	unknown	-	-	-	-	-	-	ADO67109
crov077	unknown	-	N/A	-	-	-	-	ADO67110
crov078	unknown	-	0	early	-	-	-	ADO67111
crov079	unknown	-	-	-	-	-	-	ADO67112
crov080	deoxynucleoside monophosphate kinase	-	2	early	-	1067	R512, $9x10^{-25}$	ADO67113
crov081	unknown	-	N/A	late	-	1347	-	ADO67114
crov082	ubiquitin-conjugating enzyme E2	COG5078, Ubiquitin-protein ligase [Posttranslational modification, protein turnover, chaperones] $8x10^{-9}$; PF00179, Ubiquitin-conjugating enzyme, $6x10^{-7}$	1	early	-	329	L709, $7x10^{-18}$	ADO67115
crov083	unknown	-	-	-	-	-	-	ADO67116
crov084	DNA topoisomerase	COG0550 TopA, Topoisomerase IA [DNA	2	early	-	33	**L221,**	ADO67117

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	IA	replication, recombination, and repair] $6x10^{-133}$; PF01131, DNA topoisomerase, $1x10^{-95}$; PF01751, Toprim domain, $1x10^{-29}$					**$4x10^{-129}$**	
crov085	unknown	-	N/A	late	+	-	-	ADO67118
crov086	unknown	-	N/A	-	-	-	-	ADO67119
crov087	methyltransferase	COG2226, Rossmann-fold NAD(P)(+)-binding proteins [Coenzyme metabolism] $7x10^{-8}$; PF08241, Methyltransferase domain, $3x10^{-20}$	0	early	-	-	-	ADO67120
crov088	unknown	-	72	late	+	-	-	ADO67121
crov089	unknown	-	6	-	+	-	-	ADO67122
crov090	unknown	-	-	late	+	-	-	ADO67123
crov091	unknown	-	3	early	-	-	-	ADO67124
crov092	unknown	-	6	early	-	-	-	ADO67125
crov093	unknown	-	3	early	-	-	-	ADO67126
crov094	unknown	-	6	late	+	-	-	ADO67127
crov095	unknown	-	0	early, late	-	-	-	ADO67128
crov096	unknown	-	1	early	-	-	-	ADO67129
crov097	unknown	PF08793, 2-cysteine adaptor domain, $2x10^{-4}$	-	-	-	-	-	ADO67130
crov098	unknown	-	N/A	late	+	-	-	ADO67131
crov099	unknown	-	0	early	-	-	-	ADO67132
crov100	unknown	-	-	early	-	-	-	ADO67133
crov101	unknown	-	6	-	-	-	-	ADO67134
crov102	ubiquitin-conjugating	COG5078 Ubiquitin-protein ligase [Posttranslational modification, protein	0	early	-	329	L460, $2x10^{-19}$ R795, $3x10^{-16}$	ADO67135

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	enzyme E2	turnover, chaperones] $1x10^{-47}$; PF00179, Ubiquitin-conjugating enzyme, $3x10^{-72}$					L709, $1x10^{-7}$	
crov103	unknown	PF00293, NUDIX domain, 0.017	3	early	-	-	-	ADO67136
crov104	unknown	-	N/A	early	-	-	-	ADO67137
crov105	unknown	-	unclear	-	-	-	R604, $5x10^{-7}$	ADO67138
crov106	apurinic-apyrimidinic (AP) endonuclease 1 (Apex 1)	COG0708 XthA, Exonuclease III [DNA replication, recombination, and repair] $4x10^{-62}$; PF03372, Endonuclease/Exonuclease/phosphatase family, $4x10^{-47}$	2	early	-	-	-	ADO67139
crov107	AAA+ superfamily ATPase	COG0464, SpoVK, ATPases of the AAA+ class [Posttranslational modification, protein turnover, chaperones] 9x10-12; PF00004, ATPase family associated with various cellular activities (AAA) , $7x10^{-18}$	6	-	-	780	R730, $9x10^{-53}$	ADO67140
crov108	unknown	-	N/A	early	-	-	-	ADO67141
crov109	SFII helicase / VV D6-like early transcription factor SSU	NT quarter: COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $1x10^{-4}$; PF04851, Type III restriction enzyme, res subunit, $5x10^{-7}$	6	late	+	31	**L377, 0.0 R563, $3x10^{-51}$ R350, $1x10^{-31}$ L538, $5x10^{-11}$**	ADO67142
crov110	unknown	-	2	-	+	-	-	ADO67143
crov111	unknown	PF05729, NACHT domain, $2x10^{-5}$	1	early	-	-	-	ADO67144
crov112	tRNA 2-phosphotransferase	COG1859, KptA, RNA 2'-phosphotransferase [Translation, ribosomal structure and biogenesis] $5x10^{-19}$; PF01885, RNA 2'-phosphotransferase; Tpt1 / KptA family, $9x10^{-16}$	6	early	-	-	-	ADO67145

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov113	contains eIF-2/eIF-5B domain	COG0532 InfB, Translation initiation factor 2 (IF-2; GTPase) [Translation, ribosomal structure and biogenesis] $1x10^{-61}$; PF00009, Elongation factor Tu GTP binding domain, $3x10^{-30}$	6	early	-	-	R624, $2x10^{-6}$	ADO67146
crov114	FNIP/IP22-like repeats	PF05725, FNIP, $6x10^{-13}$	6	early	-	6	L60, $5x10^{-54}$ and others	ADO67147
crov115	CPD class I photolyase	COG0415 PhrB, Deoxyribodipyrimidine photolyase [DNA replication, recombination, and repair] $1x10^{-89}$; PF03441, FAD binding domain of DNA photolyase, $1x10^{-74}$; PF00875, DNA photolyase, $3x10^{-11}$	-	-	+	-	-	ADO67148
crov116	unknown	-	6	-	-	-	-	ADO67149
crov117	unknown	-	3	early	-	-	-	ADO67150
crov118	RNA helicase	COG1643, HrpA, HrpA-like helicases [DNA replication, recombination, and repair] $1x10^{-29}$; PF00271, Helicase conserved C-terminal domain, $3x10^{-9}$	-	-	+	267	**R366, $2x10^{-105}$ L540, $5x10^{-26}$**	ADO67151
crov119	papain-like cysteine protease	PF08795, Putative papain-like cysteine peptidase (DUF1796), $5x10^{-5}$	-	-	-	-	-	ADO67152
crov120	fatty acid hydroxylase	PF04116, Fatty acid hydroxylase superfamily, $4x10^{-15}$	-	-	-	-	-	ADO67153
crov121	unknown	-	2	early	-	779	L729, $1x10^{-7}$	ADO67154
crov122	FNIP/IP22-like repeats	PF05725, FNIP, $3x10^{-9}$	N/A	-	-	-	-	ADO67155
crov123	FNIP/IP22-like repeats	PF05725, FNIP, $2x10^{-6}$	N/A	-	-	6	L60, $2x10^{-11}$ and others	ADO67156
crov124	FNIP/IP22-like	PF05725, FNIP, $1x10^{-11}$	6	early	-	6	L60, $2x10^{-22}$	ADO67157

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	repeats						and others	
crov125	phosphoglycerate mutase	COG0406, GpmB, Fructose-2,6-bisphosphatase [Carbohydrate transport and metabolism] $2x10^{-15}$; PF00300, Phosphoglycerate mutase family, $1x10^{-18}$	3	-	-	-	-	ADO67158
crov126	unknown	-	0	early	-	-	-	ADO67159
crov127	DnaJ/Hsp40	COG0484 DnaJ-class molecular chaperone with C-terminal Zn finger domain [Posttranslational modification, protein turnover, chaperones] $2x10^{-55}$; PF00226, DnaJ domain, $3x10^{-38}$; PF00684, DnaJ central domain (4 repeats), $6x10^{-10}$; PF01556, DnaJ C terminal region, $2x10^{-7}$	0	early	-	46	R260, $1x10^{-37}$ R266, $1x10^{-11}$	ADO67160
crov128	HD superfamily phosphohydrolase	COG1078, HD superfamily phosphohydrolases [General function prediction only] $7x10^{-50}$; PF01966, HD domain, $6x10^{-5}$	2	early	-	603	L394, $1x10^{-50}$	ADO67161
crov129	unknown	-	6	late	-	-	-	ADO67162
crov130	unknown	-	0	early, late	-	-	-	ADO67163
crov131	SFII helicase, VV D11-like transcription factor	COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $3x10^{-5}$; PF04851, Type III restriction enzyme, restriction subunit $5x10^{-5}$; PF08469, Nucleoside triphosphatase I C-terminal, $7x10^{-4}$; PF00271, Helicase conserved C-terminal domain, 0.006	unclear	-	+	31	**R563, $2x10^{-148}$ L377, $1x10^{-42}$ R350, $1x10^{-34}$ L538, $2x10^{-34}$**	ADO67164
crov132	unknown	-	0	late	+	-	-	ADO67165
crov133	unknown	-	unclear	-	+	-	-	ADO67166
crov134	unknown	-	12	-	-	-	-	ADO67167

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov135	unknown	-	-	-	-	-	-	ADO67168
crov136	unknown	-	6	late	+	-	-	ADO67169
crov137	insulinase-like metalloprotease	COG1025, Ptr, Secreted/periplasmic Zn-dependent peptidases, insulinase-like [Posttranslational modification, protein turnover, chaperones] $2x10^{-71}$; PF00675, Insulinase (Peptidase family M16), $3x10^{-24}$; PF05193, Peptidase M16 inactive domain, $5x10^{-9}$	0	-	+	531	L233, $4x10^{-8}$	ADO67170
crov138	glycosyltransferase	COG3306, Glycosyltransferase family 25, $7x10^{-6}$; PF01755, Glycosyltransferase family 25 (LPS biosynthesis protein), $3x10^{-5}$	-	-	-	256	R655, $4x10^{-7}$	ADO67171
crov139	5'-3' exoribonuclease	-	-	-	-	1045	**R527/R528, $2x10^{-11}$**	ADO67172
crov140	unknown	-	2	early	-	-	-	ADO67173
crov141	unknown	-	-	-	-	-	-	ADO67174
crov142	unknown	-	6	-	+	-	-	ADO67175
crov143	thiol oxidoreductase	COG5054 ERV1, Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins [Posttranslational modification, protein turnover, chaperones] $2x10^{-9}$; PF04777, Erv1 / Alr family, $2x10^{-25}$	6	-	+	52	**R596, $9x10^{-26}$** R368, $9x10^{-17}$	ADO67176
crov144	similar to conserved bacterial protein	COG3803, Uncharacterised protein conserved in bacteria [Function unknown] $6x10^{-17}$; PF06041, Bacterial protein of unknown function (DUF924), $1x10^{-19}$	2	early	-	-	-	ADO67177
crov145	unknown	-	-	late	+	-	-	ADO67178
crov146	unknown	-	-	-	-	-	-	ADO67179

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov147	ornithine decarboxylase	COG0019 LysA, Diaminopimelate decarboxylase [Amino acid transport and metabolism] $6x10^{-27}$; PF02784, Pyridoxal-dependent decarboxylase, pyridoxal binding domain, $2x10^{-22}$; PF00278, Pyridoxal-dependent decarboxylase, C-terminal sheet domain, $8x10^{-18}$	2	early	-	7	-	ADO67180
crov148	phage tail collar domain protein	COG4675 MdpB, Microcystin-dependent protein [Function unknown] $2x10^{-24}$; PF07484, Phage Tail Collar Domain, $2x10^{-19}$	6	late	+	-	-	ADO67181
crov149	photolyase	COG3046 Uncharacterised protein related to deoxyribodipyrimidine photolyase [DNA repair] $1x10^{-97}$; PF04244, Deoxyribodipyrimidine photolyase-related protein, $1x10^{-9}$; PF03441, FAD binding domain of DNA photolyase, 0.008	-	-	+	1004	R852/R853, $1x10^{-39}$ R855, $2x10^{-24}$	ADO67182
crov150	unknown	-	3	-	+	-	-	ADO67183
crov151	unknown	-	0	early	-	-	-	ADO67184
crov152	DNA topoisomerase IB	COG3569 Topoisomerase IB [DNA replication, recombination, and repair] $2x10^{-17}$; PF02919, Eukaryotic DNA topoisomerase I, DNA binding fragment, $2x10^{-72}$; PF01028, Eukaryotic DNA topoisomerase I, catalytic core, $1x10^{-72}$	unclear	-	+	-	(**R194**, putative functional homologue)	ADO67185
crov153	unknown	PF00642, Zinc finger C-x8-C-x5-C-x3-H type (and similar), $4x10^{-5}$	0	early	-	-	-	ADO67186
crov154	GIY-YIG endonuclease	COG2827 Predicted endonuclease containing a URI domain [DNA replication, recombination, and repair] $3x10^{-4}$; PF01541, GIY-YIG catalytic domain, $2x10^{-9}$	-	-	-	-	-	ADO67187
crov155	unknown	-	0	-	-	1446	L728, $8x10^{-26}$	ADO67188

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov156	unknown	-	0	early	-	-	-	ADO67189
crov157	unknown	-	0	early	-	-	-	ADO67190
crov158	unknown	-	N/A	early	-	-	-	ADO67191
crov159	unknown	-	0	early	-	-	-	ADO67192
crov160	TATA-box binding protein	no COG but cd00652, TBP_TLF, TATA box binding protein (TBP); PF00352, Transcription factor TFIID (or TATA-binding protein, TBP), 0.04	2	early	-	313	R453, $5x10^{-29}$	ADO67193
crov161	thymidine kinase	COG1435, Tdk, Thymidine kinase [Nucleotide transport and metabolism] $1x10^{-28}$; PF00265, Thymidine kinase, $1x10^{-45}$	2	early	-	319	L258, $9x10^{-15}$	ADO67194
crov162	contains eIF-2alpha domain	COG1093, SUI2, Translation initiation factor 2, alpha subunit (eIF-2alpha) [Translation, ribosomal structure and biogenesis] $4x10^{-8}$;	6	early	-	-	-	ADO67195
crov163	Holliday junction resolvase	PF09159, Mitochondrial resolvase Ydc2, catalytic, $5x10^{-4}$	-	-	-	278	L451, $2x10^{-35}$	ADO67196
crov164	transcription factor	-	-	-	-	1164	R450, $4x10^{-12}$	ADO67197
crov165	unknown	-	-	-	+	-	-	ADO67198
crov166	unknown	-	6	-	-	1423	R449, $2x10^{-28}$	ADO67199
crov167	unknown	-	-	late	+	-	-	ADO67200
crov168	unknown	-	N/A	-	-	-	R447, $4x10^{-6}$	ADO67201
crov169	papatin-like phospholipase	COG1752 RssA, Predicted esterase of the alpha-beta hydrolase superfamily [General function prediction only] $7x10^{-19}$; PF01734, Patatin-like phospholipase, $8x10^{-17}$	-	-	-	245	L446, $5x10^{-26}$ R571, $1x10^{-19}$ L620, $8x10^{-7}$	ADO67202
crov170	contains DnaJ/Hsp40	COG0484 DnaJ-class molecular chaperone with C-terminal Zn finger domain	6	late	-	46	R260, $9x10^{-11}$	ADO67203

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	domain	[Posttranslational modification, protein turnover, chaperones] $1x10^{-12}$; PF00226, DnaJ domain, $2x10^{-19}$					R445, $6x10^{-6}$	
crov171	ADP-ribosyl glycohydrolase	COG1397 DraG, ADP-ribosylglycohydrolase [Posttranslational modification, protein turnover, chaperones] $7x10^{-7}$; PF03747, ADP-ribosylglycohydrolase, $4x10^{-6}$	-	early	+	3	L444, $4x10^{-29}$	ADO67204
crov172	protein disulfide isomerase	COG0526, TrxA, Thiol-disulfide isomerase and thioredoxins [Posttranslational modification, protein turnover, chaperones / Energy production and conversion] 0.006; PF00085, Thioredoxin, $8x10^{-9}$	-	late	+	-	**R443, $1x10^{-5}$**	ADO67205
crov173	unknown	-	-	-	+	-	-	ADO67206
crov174	unknown	-	6	late	+	-	-	ADO67207
crov175	unknown	-	-	late	+	-	-	ADO67208
crov176	capsid protein 4	PF04451, Large eukaryotic DNA virus major capsid protein, $1x10^{-8}$	-	-	+	22	R441, $3x10^{-64}$ **L425, $9x10^{-35}$** R439, $3x10^{-16}$ R440, $3x10^{-15}$	ADO67209
crov177	unknown	-	0	early	-	-	-	ADO67210
crov178	unknown	-	-	-	-	22	R439, $4x10^{-13}$ **L425, $4x10^{-12}$** R441, $2x10^{-9}$ R440, $3x10^{-9}$	ADO67211
crov179	unknown	-	-	-	+	-	-	ADO67212
crov180	unknown	-	N/A	late	-	-	-	ADO67213
crov181	unknown	-	-	early	-	-	-	ADO67214
crov182	dsRNA-specific	COG0571 dsRNA-specific ribonuclease	1	early	-	1354	R343, $2x10^{-39}$	ADO67215

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	ribonuclease III	[Transcription] $3x10^{-34}$; PF00636, RNase3 domain, $8x10^{-35}$; PF00035, Double-stranded RNA binding motif, $6x10^{-8}$						
crov183	unknown	-	6	-	-	-	-	ADO67216
crov184	cysteine protease	PF02902, Ulp1 protease family, C-terminal catalytic domain, $2x10^{-5}$	6	-	+	246	**R355, $2x10^{-52}$**	ADO67217
crov185	unknown	-	6	-	+	1131	**L454, $3x10^{-24}$**	ADO67218
crov186	unknown	-	-	late	+	628	**R457, $4x10^{-13}$**	ADO67219
crov187	unknown	-	12	late	+	-	-	ADO67220
crov188	contains ankyrin repeats	PF00023, Ankyrin repeat, 0.003	-	-	-	-	-	ADO67221
crov189	unknown	-	0	early	-	-	-	ADO67222
crov190	contains SMC N-terminal domain	COG4942 Membrane-bound metallopeptidase [Cell division and chromosome partitioning] $9x10^{-6}$	3	early	-	-	-	ADO67223
crov191	unknown	-	N/A	-	-	-	-	ADO67224
crov192	unknown	-	-	-	-	-	-	ADO67225
crov193	unknown	-	N/A	early	-	-	-	ADO67226
crov194	unknown	-	12	late	+	1083	R468, $4x10^{-6}$	ADO67227
crov195	replication factor C small subunit	COG2812, DnaX, DNA polymerase III, gamma/tau subunits [DNA replication, recombination, and repair] $2x10^{-24}$; PF00004, ATPase family associated with various cellular activities (AAA), $8x10^{-17}$	3	early	-	1351	L510, $2x10^{-48}$ L499, $4x10^{-44}$ R395, $2x10^{-21}$ L478, $2x10^{-18}$	ADO67228
crov196	replication factor C small subunit	COG0470, HolB, ATPase involved in DNA replication [DNA replication, recombination, and repair] $8x10^{-26}$; PF00004, ATPase family	unclear	-	-	1351	L510, $2x10^{-43}$ L499, $1x10^{-37}$ R395, $9x10^{-30}$	ADO67229

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		associated with various cellular activities (AAA), $7x10^{-7}$					L478, $2x10^{-10}$	
crov197	unknown	-	N/A	-	-	-	-	ADO67230
crov198	ubiquitin-conjugating enzyme E2	COG5078, Ubiquitin-protein ligase [Posttranslational modification, protein turnover, chaperones] $2x10^{-15}$; PF00179, Ubiquitin-conjugating enzyme, $5x10^{-30}$	2	early	-	329	R521, $5x10^{-57}$	ADO67231
crov199	unknown	-	6	-	+	-	-	ADO67232
crov200	exonuclease	PF01612, 3'-5' exonuclease, 0.006	-	-	+	671	**L533, $5x10^{-17}$**	ADO67233
crov201	Rpb10, DNA-directed RNA polymerase II subunit 10	COG1644, RPB10, DNA-directed RNA polymerase, subunit N (RpoN/RPB10) [Transcription] 0.002; PF01194, RNA polymerases N / 8 kDa subunit, $3x10^{-5}$	N/A	-	+	-	-	ADO67234
crov202	unknown	-	-	-	+	-	-	ADO67235
crov203	contains RING finger domain	COG5540, RING-finger-containing ubiquitin ligase [Posttranslational modification, protein turnover, chaperones] $3x10^{-9}$	0	-	-	-	-	ADO67236
crov204	unknown	-	-	-	-	-	-	ADO67237
crov205	unknown	-	-	-	-	-	-	ADO67238
crov206	lambda-type exonuclease	COG5377, viral recombinase [DNA replication, recombination, and repair] 0.01; PF09588, YqaJ viral recombinase family, $2x10^{-35}$	2	early	-	1192	R354, $4x10^{-56}$	ADO67239
crov207	unknown	-	-	-	-	-	-	ADO67240
crov208	unknown	-	2	early	-	-	-	ADO67241
crov209	unknown	-	0	early	-	-	-	ADO67242
crov210	contains DnaJ/Hsp40 domain	COG0484, DnaJ, [Posttranslational modification, protein turnover, chaperones]	-	-	-	-	-	ADO67243

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		$6x10^{-5}$; PF00226, DnaJ domain, 0.001						
crov211	unknown	-	N/A	-	-	-	-	ADO67244
crov212	mRNA capping enzyme	PF03291, mRNA capping enzyme, $2x10^{-11}$; PF02940, mRNA capping enzyme, beta chain, $3x10^{-5}$; PF01331, mRNA capping enzyme, catalytic domain, $3x10^{-4}$	3	-	+	1117	**R382, $1x10^{-127}$**	ADO67245
crov213	unknown	-	-	late	+	-	-	ADO67246
crov214	unknown	-	6	early	-	-	-	ADO67247
crov215	unknown	-	6	-	+	648	**L492, $2x10^{-10}$**	ADO67248
crov216	nucleotidyl transferase	COG1208, Nucleoside-diphosphate-sugar pyrophosphorylase involved in lipopolysaccharide biosynthesis/translation initiation factor 2B, gamma/epsilon subunits (eIF-2Bgamma/eIF-2Bepsilon) [Cell envelope biogenesis, outer membrane / Translation, ribosomal structure and biogenesis] $3x10^{-6}$; PF00483, Nucleotidyl transferase, 0.008	-	-	-	235	-	ADO67249
crov217	unknown	-	6	late	+	-	-	ADO67250
crov218	NAD-dependent epimerase/ dehydratase	COG0451, Nucleoside-diphosphate-sugar epimerases [Cell envelope biogenesis, outer membrane / Carbohydrate transport and metabolism] $2x10^{-48}$; PF01370, NAD dependent epimerase.dehydratase family, $5x10^{-53}$	6	-	-	1049	R141, $1x10^{-21}$	ADO67251
crov219	proliferating cell nuclear antigen (PCNA)	COG0592, DnaN, DNA polymerase sliding clamp subunit (PCNA homolog) [DNA replication, recombination, and repair] 0.003; PF00705, Proliferating cell nuclear antigen, N-terminal domain, $7x10^{-6}$; PF02747, Proliferating cell nuclear antigen, C-terminal	2	early	-	241	R493, $2x10^{-20}$	ADO67252

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		domain, $6x10^{-6}$						
crov220	eIF-4E	COG5053, CDC33, Translation initiation factor 4E (eIF-4E) [Translation, ribosomal structure and biogenesis] $8x10^{-6}$; PF01652, Eukaryotic initiation factor 4E, $2x10^{-5}$	12	-	-	651	L496, $3x10^{-8}$	ADO67253
crov221	unknown	-	-	-	-	-	-	ADO67254
crov222	unknown	-	0	-	+	-	-	ADO67255
crov223	unknown	-	6	-	-	-	-	ADO67256
crov224	Rpb2, DNA-directed RNA polymerase II subunit 2, intein-containing	COG0085 RpoB, DNA-directed RNA polymerase, beta subunit/140 kD subunit [Transcription] $9x10^{-103}$; PF04563, RNA polymerase beta subunit, $2x10^{-9}$; PF04561, RNA polymerase Rpb2, domain 2, $2x10^{-15}$; PF04565, RNA polymerase Rpb2, domain 3, $5x10^{-16}$; PF04566, RNA polymerase Rpb2, domain 4, $9x10^{-14}$; PF00562, RNA polymerase Rpb2, domain 6, $8x10^{-95}$; PF04560, RNA polymerase Rpb2, $5x10^{-31}$	2	early	+	271	**L244, 0.0**	ADO67257
crov225	unknown	-	-	-	-	-	-	ADO67258
crov226	unknown	-	6	-	-	647	L491, $6x10^{-10}$	ADO67259
crov227	unknown	-	-	-	-	-	R521, $5x10^{-7}$	ADO67260
crov228	unknown	-	6	late	+	645	**R489, $1x10^{-18}$**	ADO67261
crov229	contains ankyrin repeats	PF00023, Ankyrin repeat, $1x10^{-6}$	-	-	-	-	-	ADO67262
crov230	transcription initiation factor TFIIB	COG1405, SUA7, Transcription initiation factor TFIIB, Brf1 subunit/Transcription initiation factor TFIIB [Transcription] $7x10^{-30}$; PF00382, Transcription factor TFIIB repeat, $8x10^{-8}$	3	early	-	1127	L250, $5x10^{-51}$	ADO67263

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov231	unknown	-	-	-	-	-	L371, $4x10^{-7}$	ADO67264
crov232	ABC-type amino acid transporter	COG0834, HisJ, ABC-type amino acid transport/signal transduction systems, periplasmic component/domain [Amino acid transport and metabolism / Signal transduction mechanisms] $4x10^{-12}$; PF00497, Bacterial extracellular solute-binding proteins, family 3, $1x10^{-23}$	0	early	+	-	-	ADO67265
crov233	unknown	-	0	early	-	-	-	ADO67266
crov234	unknown	-	N/A	early	-	-	-	ADO67267
crov235	unknown	-	N/A	early	-	-	-	ADO67268
crov236	unknown	-	1	early	-	-	-	ADO67269
crov237	Lon protease	COG0466 Lon, ATP-dependent Lon protease, bacterial type [Posttranslational modification, protein turnover, chaperones] $3x10^{-86}$; PF05362, Lon protease (S16) C-terminal proteolytic domain, $8x10^{-33}$; PF00004, ATPase family associated with various cellular activities (AAA), $4x10^{-25}$	6	-	-	228	L251, $9x10^{-177}$	ADO67270
crov238	unknown	-	6	late	-	329	R521, $1x10^{-12}$ R252, $5x10^{-11}$	ADO67271
crov239	unknown	-	6	-	-	-	-	ADO67272
crov240	unknown	-	N/A	late	+	-	-	ADO67273
crov241	DnaK/Hsp70	COG0443 DnaK, Molecular chaperone [Posttranslational modification, protein turnover, chaperones] $9x10^{-38}$; PF00012, Hsp70 protein, $2x10^{-9}$	-	-	-	73	L393, $5x10^{-31}$ L254, $9x10^{-30}$	ADO67274
crov242	UDP-Glc 6-	COG1004, Ugd, Predicted UDP-glucose 6-	0	-	-	332	-	ADO67275

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	dehydrogenase	dehydrogenase [Cell envelope biogenesis, outer membrane] $2x10^{-28}$; PF03721, UDP-glucose/GDP-mannose dehydrogenase family, NAD binding domain, $2x10^{-6}$; PF00984, UDP-glucose/GDP-mannose dehydrogenase family, central domain, $2x10^{-5}$						
crov243	unknown	-	-	-	-	-	-	ADO67276
crov244	unknown	-	3	-	-	-	-	ADO67277
crov245	unknown	-	N/A	-	-	-	-	ADO67278
crov246	unknown	-	-	-	-	-	-	ADO67279
crov247	glycosyltransferase	COG1215, Glycosyltransferases, probably involved in cell wall biogenesis [Cell envelope biogenesis, outer membrane] 0.006; PF00535, Glycosyl transferase family 2, $6x10^{-14}$	-	-	-	-	-	ADO67280
crov248	unknown	-	N/A	-	-	-	-	ADO67281
crov249	unknown	-	N/A	-	-	-	-	ADO67282
crov250	unknown	-	-	-	-	-	-	ADO67283
crov251	unknown	-	N/A	-	-	-	-	ADO67284
crov252	unknown	-	3	-	-	-	-	ADO67285
crov253	unknown	-	-	-	-	-	-	ADO67286
crov254	unknown	-	6	-	-	-	-	ADO67287
crov255	unknown	-	-	-	-	-	-	ADO67288
crov256	unknown	-	6	-	-	-	-	ADO67289
crov257	unknown	-	N/A	-	-	-	-	ADO67290
crov258	unknown	-	N/A	-	-	-	-	ADO67291

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov259	unknown	-	-	-	-	-	-	ADO67292
crov260	exostosin family protein	PF03016, Exostosin family, $4x10^{-13}$	-	-	-	-	-	ADO67293
crov261	glycosyltransferase/S-adenosylmethionine-dependent methyltransferase	COG2227, UbiG, 2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-1,4-benzoquinol methylase [Coenzyme metabolism] $5x10^{-6}$ / COG0438, RfaG, Glycosyltransferase [Cell envelope biogenesis, outer membrane] $4x10^{-4}$; PF08241, Methyltransferase domain, $6x10^{-16}$	-	-	-	-	-	ADO67294
crov262	unknown	-	12	-	-	-	-	ADO67295
crov263	unknown	-	-	-	-	-	-	ADO67296
crov264	unknown	-	-	-	-	-	-	ADO67297
crov265	bifunctional KDO 8-P phosphatase/ arabinose 5-P isomerase	NT 3rd: COG1778 Low specificity phosphatase (HAD superfamily) [General function prediction only] $1x10^{-33}$; middle 3rd: COG0794 GutQ, Predicted sugar phosphate isomerase involved in capsule formation [Cell envelope biogenesis, outer membrane] $2x10^{-21}$; PF08282, haloacid dehalogenase-like hydrolase, $8x10^{-6}$; PF01380, SIS domain, $9x10^{-11}$	-	-	-	-	-	ADO67298
crov266	bifunctional N-acylneuraminate cytidylyltransferase/ demethyl-menaquinone methyltransferase	COG0684 MenG, Demethylmenaquinone methyltransferase [Coenzyme metabolism] $2x10^{-14}$, COG1083 NeuA, CMP-N-acetylneuraminic acid synthetase [Cell envelope biogenesis, outer membrane] $8x10^{-14}$; PF02348, Cytidylyltransferase, $5x10^{-13}$	-	-	-	-	-	ADO67299
crov267	bifunctional KDO 8-P synthase/dTDP-6-deoxy-L-hexose 3-O-	NT half: COG2877 KdsA, 3-deoxy-D-manno-octulosonic acid (KDO) 8-phosphate synthase [Cell envelope biogenesis, outer membrane]	6	-	-	-	-	ADO67300

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	methyltransferase	$8x10^{-82}$; PF00793, DAHP synthetase I family, $1x10^{-68}$						
crov268	unknown	-	-	-	-	-	-	ADO67301
crov269	unknown	PF01697, Domain of unknown function, $1x10^{-4}$	3	-	-	-	-	ADO67302
crov270	unknown	-	-	-	-	-	-	ADO67303
crov271	unknown	-	6	-	-	-	-	ADO67304
crov272	unknown	-	6	-	-	-	-	ADO67305
crov273	glycosyltransferase	COG0438, RfaG, Glycosyltransferase [Cell envelope biogenesis, outer membrane] $2x10^{-5}$; PF00534, Glycosyl transferases group 1, $2x10^{-6}$	-	-	-	-	-	ADO67306
crov274	unknown	-	-	-	-	-	-	ADO67307
crov275	unknown	-	-	-	-	-	-	ADO67308
crov276	unknown	-	6	late	-	-	-	ADO67309
crov277	contains RING finger domain	COG5219, Uncharacterised conserved protein, contains RING Zn-finger [General function prediction only] $3x10^{-7}$; PF00097, Zinc finger, C3HC4 type (RING finger), $1x10^{-8}$	N/A	early	-	-	-	ADO67310
crov278	unknown	-	6	-	-	-	-	ADO67311
crov279	ATPase	PF00004, ATPase family associated with various cellular activities (AAA), $8x10^{-12}$	6	-	-	780	R730, $5x10^{-51}$	ADO67312
crov280	unknown	-	-	-	-	-	-	ADO67313
crov281	ubiquitin-conjugating enzyme E2	COG5078, Ubiquitin-protein ligase [Posttranslational modification, protein turnover, chaperones] $8x10^{-40}$; PF00179, Ubiquitin-conjugating enzyme, $1x10^{-56}$	0, cross-hybridization with host mRNA	early, late	-	329	L460, $1x10^{-19}$ R795, $1x10^{-14}$ R521, $8x10^{-10}$ L709, $4x10^{-6}$	ADO67314

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov282	unknown	-	-	-	-	-		ADO67315
crov283	SFII helicase/VV D11-like transcription factor	COG0553, HepA, Superfamily II DNA/RNA helicases, SNF2 family [Transcription / DNA replication, recombination, and repair] $7x10^{-5}$ / COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $8x10^{-5}$; PF00271, Helicase conserved C-terminal domain, $2x10^{-10}$; PF00176, SNF2 family N-terminal domain, $2x10^{-5}$	3	late	+	31	**R350, 0.0 R563, $3x10^{-36}$ L538, $1x10^{-25}$ L377, $5x10^{-13}$**	ADO67316
crov284	unknown	-	0	early	-	-	-	ADO67317
crov285	unknown	-	0	early	-	-	-	ADO67318
crov286	polyadenylate polymerase catalytic subunit	-	0, cross-hybridization with host mRNA	-	+	1165	**R341, $2x10^{-56}$**	ADO67319
crov287	unknown	-	N/A	early	-	-	-	ADO67320
crov288	unknown	-	6	-	+	-	-	ADO67321
crov289	unknown	-	N/A	-	-	-	-	ADO67322
crov290	unknown	-	6	late	-	-	-	ADO67323
crov291	metal-dependent hydrolase	-	-	late	-	1120	R325, $5x10^{-43}$	ADO67324
crov292	VV A7-like early transcription factor LSU	-	1	-	+	261	**R326/R327, 0.0**	ADO67325
crov293	unknown	-	N/A	late	-	-	-	ADO67326
crov294	unknown	-	-	late	+	-	-	ADO67327

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov295	unknown	-	3	-	+	-	-	ADO67328
crov296	unknown	-	-	late	+	-	-	ADO67329
crov297	unknown	-	6	early	-	-	-	ADO67330
crov298	unknown	-	0	early	-	573	R338, $1x10^{-32}$	ADO67331
crov299	transcription elongation factor TFIIS	CT third: COG1594, RPB9, DNA-directed RNA polymerase, subunit M/Transcription elongation factor TFIIS [Transcription] $2x10^{-9}$; PF01096, Transcription factor S-II (TFIIS), $3x10^{-14}$	2	early	+	272	R339, $2x10^{-10}$	ADO67332
crov299a	unknown	-	N/A	-	+	-	-	ADO67333
crov300	unknown	-	N/A	early	-	-	-	ADO67334
crov301	contains RNA recognition motif	PF00076, RNA recognition motif. (also known as RRM, RBD, or RNP domain), $1x10^{-4}$	-	-	-	-	-	ADO67335
crov302	AAA family ATPase	COG0464, SpoVK, ATPases of the AAA+ class [Posttranslational modification, protein turnover, chaperones] $5x10^{-8}$; COG1222, RPT1, ATP-dependent 26S proteasome regulatory subunit [Posttranslational modification, protein turnover, chaperones] $7x10^{-8}$; PF00004, ATPase family associated with various cellular activities (AAA), $7x10^{-12}$	0	early	-	1	L572/L573, $8x10^{-27}$ R776, $3x10^{-8}$	ADO67336
crov303	formamidopyrimidine-DNA glycosylase	COG0266, Nei, Formamidopyrimidine-DNA glycosylase [DNA replication, recombination, and repair] $6x10^{-29}$; PF01149, Formamidopyrimidine-DNA glycosylase N-terminal domain, $2x10^{-18}$; PF06831, Formamidopyrimidine-DNA glycosylase H2TH domain, $1x10^{-15}$	6	-	+	562	L315, $2x10^{-40}$	ADO67337
crov304	5'-3' exonuclease	COG5049 XRN1, 5'-3' exonuclease [DNA	2	early	-	1045		

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		replication, recombination, and repair / Cell division and chromosome partitioning / Translation] $2x10^{-46}$; PF03159, XRN 5'-3' exonuclease N-terminus, $2x10^{-57}$					**R527/R528, $2x10^{-14}$**	ADO67338
crov305	unknown	-	0	-	-	-	-	ADO67339
crov306	unknown	-	0	-	-	-	-	ADO67340
crov307	unknown	-	N/A	-	-	-	-	ADO67341
crov308	tubulin-tyrosine ligase family protein	-	0	-	-	-	-	ADO67342
crov309	Ser/Thr protein kinase	PF00069, Protein kinase domain, $3x10^{-5}$	0	-	+	295	**R400, $1x10^{-82}$ R402, $3x10^{-8}$ R301, $1x10^{-7}$**	ADO67343
crov310	unknown	-	-	-	+	-	-	ADO67344
crov311	unknown	-	-	-	+	637	**R472, $2x10^{-98}$**	ADO67345
crov312	unknown	-	6	late	+	-	-	ADO67346
crov313	phosphoesterase	PF00149, Calcineurin-like phosphoesterase, $4x10^{-11}$	2	-	+	995	**R398, $1x10^{-39}$**	ADO67347
crov314	unknown	-	-	-	-	-	-	ADO67348
crov315	unknown	-	0	-	-	-	-	ADO67349
crov316	VV A18-like helicase	COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $5x10^{-39}$; PF054851, Type III restriction enzyme, res subunit, $1x10^{-21}$	6	early	-	76	L396, $1x10^{-73}$ L173, $1x10^{-6}$	ADO67350
crov317	unknown	-	6	-	-	-	-	ADO67351
crov318	unknown	-	N/A	-	+	642	**L485, $1x10^{-09}$**	ADO67352

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov319	contains ankyrin repeats	COG0666, Arp, FOG: Ankyrin repeat [General function prediction only] 0.001; PF00023, Ankyrin repeat, $7x10^{-10}$	-	-	+	-	**L484, $3x10^{-8}$** L289, $9x10^{-8}$	ADO67353
crov320	unknown	-	6	late	+	-	-	ADO67354
crov321	capsid protein 3	PF04451, Large eukaryotic DNA virus major capsid protein, $5x10^{-6}$	-	late	+	22	R439, $1x10^{-23}$ R441, $2x10^{-18}$ **L425, $2x10^{-13}$** R440, $2x10^{-10}$	ADO67355
crov322	unknown	-	-	-	-	-	-	ADO67356
crov323	unknown	-	0	early	-	-		ADO67357
crov324	unknown	-	-	-	-	-	-	ADO67358
crov325	DNA topoisomerase IIA intein-containing	COG0187 Type IIA topoisomerase [DNA replication, recombination, and repair] $4x10^{-52}$ / COG0188, GyrA, Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit [DNA replication, recombination, and repair] $5x10^{-49}$; PF00521, DNA gyrase/topoisomerase IV, subunit A, $1x10^{-45}$; PF00204, DNA gyrase B, $7x10^{-14}$	2	early	-	37	R480, 0.0	ADO67359
crov326	unknown	-	-	-	+	641	R481, $1x10^{-19}$	ADO67360
crov327	unknown	PF03748, Flagellar basal body-associated protein FliL, $2x10^{-4}$	6	late	+	-	-	ADO67361
crov328	unknown	-	0	early	-	-	-	ADO67362
crov329	replication factor C small subunit	COG0470, HolB, ATPase involved in DNA replication [DNA replication, recombination, and repair] $8x10^{-4}$;	-	early	-	1351	L478, $4x10^{-17}$ L510, $1x10^{-6}$ L499, $5x10^{-6}$	ADO67363
crov330	unknown	-	6	-	+	608	**R403, $1x10^{-64}$ R402, $6x10^{-12}$**	ADO67364

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov331	unknown	-	-	-	+	1129	R409, $1x10^{-18}$	ADO67365
crov332	major core protein	PF03292, Poxvirus P4B major core protein, 0.02	6	late	+	1162	**L410, $2x10^{-69}$**	ADO67366
crov333	unknown	-	-	early	-	-	R411, $3x10^{-23}$	ADO67367
crov334	unknown	-	-	-	-	-	-	ADO67368
crov335	unknown	-	-	late	-	-	-	ADO67369
crov336	unknown	-	6	late	+	1216	**L417, $1x10^{-15}$**	ADO67370
crov337	unknown	-	N/A	-	-	-	-	ADO67371
crov338	VV A32-like packaging ATPase	PF04665, Poxvirus A32 protein, 0.007	6	-	-	249	L437, $2x10^{-51}$	ADO67372
crov339	unknown	-	3	-	-	158	R435, $3x10^{-8}$	ADO67373
crov340	unknown	-	N/A	-	+	-	-	ADO67374
crov341	Poxvirus VLTF3-like	PF04947, Poxvirus Late Transcription Factor VLTF3 like, $2x10^{-44}$	unclear	-	-	262	R429, $2x10^{-33}$ R430, $2x10^{-13}$	ADO67375
crov342	major capsid protein	PF04451, Large eukaryotic DNA virus major capsid protein, $7x10^{-31}$	6	late	+	22	**L425, $3x10^{-90}$** R441, $2x10^{-84}$ R439, $4x10^{-46}$ R440, $1x10^{-22}$	ADO67376
crov343	cysteine protease	PF08795, Putative papain-like cysteine peptidase (DUF1796), $6x10^{-7}$	6	early	-	609	R408, $6x10^{-10}$	ADO67377
crov344	unknown	-	3	early	-	-	-	ADO67378
crov345	unknown	-	-	-	-	67	-	ADO67379
crov346	contains hydroxylase domain	COG3128, PiuC, Uncharacterised iron-regulated protein [Function unknown] $8x10^{-7}$; PF03171, 2OG-Fe(II) oxygenase superfamily, $6x10^{-4}$	0	-	-	-	-	ADO67380

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov347	DNA N6-adenine methyltransferase	COG0338, Dam, Site-specific DNA methylase [DNA replication, recombination, and repair] $1x10^{-18}$; PF02086, D12 class N6 adenine-specific DNA methyltransferase, $2x10^{-30}$	24	-	-	232	-	ADO67381
crov348	unknown	-	N/A	-	-	-	-	ADO67382
crov349	prolyl 4-hydroxylase alpha subunit	PF03141, 2OG-Fe(II) oxygenase superfamily, $1x10^{-9}$	-	-	-	1166	**L593, $6x10^{-30}$**	ADO67383
crov350	ubiquitin	COG5272, UBI4, Ubiquitin [Posttranslational modification, protein turnover, chaperones] $4x10^{-4}$; PF00240, Ubiquitin family, $3x10^{-36}$	N/A	early	-	333	-	ADO67384
crov351	unknown	-	3	-	+	518	L202, $6x10^{-20}$	ADO67385
crov352	unknown	-	-	late	-	-	-	ADO67386
crov353	unknown	-	-	-	-	519	R203, $6x10^{-9}$	ADO67387
crov354	deoxycytidylate deaminase	COG2131, RibD, Cytidine deaminase [Coenzyme metabolism] $1x10^{-31}$; PF00383, Cytidine and deoxycytidylate deaminase zinc-binding region, $8x10^{-33}$	-	-	-	1064	-	ADO67388
crov355	unknown	-	N/A	-	-	-	-	ADO67389
crov356	transposase	COG0675, Transposase and inactivated derivatives [DNA replication, recombination, and repair] $2x10^{-6}$; PF07282, Putative transposase DNA-binding domain, $3x10^{-19}$	6	-	-	326	L770, $6x10^{-59}$ R854, $1x10^{-58}$ L79, $7x10^{-57}$ R104, $6x10^{-34}$ R186, $3x10^{-33}$	ADO67390
crov357	resolvase	COG2452, Predicted site-specific integrase-resolvase [DNA replication, recombination, and repair] $2x10^{-22}$; PF00239, Resolvase, N terminal domain, $1x10^{-13}$; PF00376, MerR family regulatory protein, $3x10^{-8}$	unclear	-	-	1175	R771, $5x10^{-56}$ R80, $2x10^{-55}$ L103, $4x10^{-42}$	ADO67391

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov358	unknown	-	0	late	+	-	-	ADO67392
crov359	similar to fibronectin domain-containing protein	PF10102, Domain of unknown function (DUF2341), $3x10^{-15}$; PF05887, Procyclic acidic repetitive protein (PARP), $3x10^{-8}$	6	-	-	-	-	ADO67393
crov360	unknown	-	2	-	-	-	-	ADO67394
crov361	unknown	-	6	late	+	-	-	ADO67395
crov362	unknown	-	N/A	early	-	-	-	ADO67396
crov363	unknown	-	0, cross-hybridization with host mRNA	late	+	-	-	ADO67397
crov364	unknown	-	-	-	-	-	-	ADO67398
crov365	unknown	-	6	late	+	-	-	ADO67399
crov366	unknown	-	N/A	late	+	-	-	ADO67400
crov367	unknown	-	6	-	-	-	-	ADO67401
crov368	Rpb1, DNA-directed RNA polymerase II subunit 1	COG0086, RpoC, DNA-directed RNA polymerase, beta' subunit/160 kD subunit [Transcription] $3x10^{-127}$; PF04997, RNA polymerase Rpb1, domain 1, $5x10^{-61}$; PF00623, RNA polymerase Rpb1, domain 2, $5x10^{-80}$; PF04983, RNA polymerase Rpb1, domain 3, $3x10^{-32}$; PF05000, RNA polymerase Rpb1, domain 4, $7x10^{-19}$; PF04998, RNA polymerase Rpb1, domain 5, $1x10^{-70}$	2	early	+	274	**R501, 0.0**	ADO67402
crov369	unknown	PF00982, Glycosyltransferase family 20, 0.008	-	-	-	-	-	ADO67403
crov370	SFII helicase/eIF-4AIII	COG0513 SrmB, Superfamily II DNA and RNA helicases [DNA replication, recombination, and repair / Transcription / Translation, ribosomal	6	-	-	25	R458, $1x10^{-46}$ R290, $9x10^{-13}$	ADO67404

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		structure and biogenesis] $8x10^{-83}$; PF00270, DEAD/DEAH box helicase, $2x10^{-48}$; PF00271, Helicase conserved C-terminal domain, $8x10^{-37}$						
crov371	unknown	-	N/A	-	-	-	-	ADO67405
crov372	unknown	-	0	early	-	657	R505, $8x10^{-12}$	ADO67406
crov373	unknown	-	-	-	-	1137	L507, $4x10^{-11}$	ADO67407
crov374	unknown	-	1	early	-	-	-	ADO67408
crov375	contains DnaJ/Hsp40 domain	NT 3rd: COG0484, DnaJ, DnaJ-class molecular chaperone with C-terminal Zn finger domain [Posttranslational modification, protein turnover, chaperones] $2x10^{-12}$; PF00226, DnaJ domain, $2x10^{-24}$	-	-	-	46	R260, $9x10^{-10}$ R266, $7x10^{-7}$	ADO67409
crov376	unknown	-	-	-	-	-	-	ADO67410
crov377	unknown	-	N/A	-	-	-	-	ADO67411
crov378	unknown	-	-	-	-	-	-	ADO67412
crov379	protein disulfide isomerase	COG0526, TrxA, Thiol-disulfide isomerase and thioredoxins [Posttranslational modification, protein turnover, chaperones / Energy production and conversion] 0.004; PF00085, Thioredoxin, $6x10^{-19}$	-	late	+	317	**R443, $5x10^{-8}$**	ADO67413
crov380	unknown	-	N/A	-	-	-	-	ADO67414
crov381	unknown	-	N/A	-	-	-	-	ADO67415
crov382	SFII RNA helicase	COG4581, Superfamily II RNA helicase [DNA replication, recombination, and repair] $2x10^{-104}$; PF00270; PF08148, DEAD/DEAH box helicase, $5x10^{-22}$; DSHCT (NUC185) domain, $2x10^{-9}$	3	early	-	30	-	ADO67416

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov383	type I restriction modification enzyme, M and S domains	COG0286, HsdM, Type I restriction-modification system methyltransferase subunit [Defence mechanisms] $2x10^{-7}$; PF01420; PF02384, Type I restriction modification DNA specificity domain, $3x10^{-16}$; N-6 DNA Methylase, $4x10^{-6}$	12	early	-	-	-	ADO67417
crov384	unknown	PF04851, Type III restriction enzyme; res subunit, $5x10^{-5}$	6	early	-	-	-	ADO67418
crov385	unknown	-	N/A	early	-	-	-	ADO67419
crov386	ubiquitin carboxyl-terminal hydrolase	COG5533, UBP5, Ubiquitin C-terminal hydrolase [Posttranslational modification, protein turnover, chaperones] $5x10^{-15}$; PF00443, Ubiquitin carboxyl-terminal hydrolase, $1x10^{-24}$	-	early	-	-	R319, $8x10^{-11}$	ADO67420
crov387	exoribonuclease	COG0557, VacB, Exoribonuclease R [Transcription] $1x10^{-12}$; PF00773, RNB domain, $4x10^{-6}$	-	early	-	-	-	ADO67421
crov388	unknown	-	0	early	-	-	-	ADO67422
crov389	SFII RNA helicase	COG1204, Superfamily II helicase [General function prediction only] $2x10^{-10}$ / COG4581, Superfamily II RNA helicase [DNA replication, recombination, and repair] $1x10^{-9}$; PF00270, DEAD/DEAH box helicase, $2x10^{-8}$	0	early	-	-	-	ADO67423
crov390	unknown	-	6	-	+	-	R554, $1x10^{-8}$ L138, $7x10^{-6}$	ADO67424
crov391	ELP3 family histone acetyltransferase	COG1243 ELP3, Histone acetyltransferase [Transcription / Chromatin structure and dynamics] $2x10^{-46}$; PF04055, Radical SAM superfamily, $1x10^{-12}$; PF00583, Acetyltransferase (GNAT) family, $1x10^{-7}$	0	early	-	-	-	ADO67425

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov392	N-acetyltransferase	PF00583, Acetyltransferase (GNAT) family, $7x10^{-18}$	-	late	-	63	L316, $2x10^{-19}$	ADO67426
crov393	Calcineurin-like phosphoesterase	PF00149, Calcineurin-like phosphoesterase, $1x10^{-4}$	2	early	-	350	L509, $9x10^{-18}$	ADO67427
crov394	unknown	-	3	early	-	-	-	ADO67428
crov395	unknown	-	6	-	-	-	-	ADO67429
crov396	DNA methyltransferase	COG0286, HsdM, Type I restriction-modification system methyltransferase subunit [Defence mechanisms] $4x10^{-4}$; PF07669, Eco57I restriction endonuclease, $1x10^{-4}$	3	-	-	-	-	ADO67430
crov397	unknown	-	N/A	-	-	-	-	ADO67431
crov398	capsid protein 2	PF04451, Large eukaryotic DNA virus major capsid protein, $2x10^{-13}$	6	late	+	22	**L425, $4x10^{-28}$** R439, $2x10^{-27}$ R441, $3x10^{-26}$ R440, $2x10^{-6}$	ADO67432
crov399	unknown	-	6	-	+	-	-	ADO67433
crov400	unknown	-	-	late	+	-	-	ADO67434
crov401	unknown	-	N/A	-	+	-	-	ADO67435
crov402	SFII helicase	COG0553, HepA, Superfamily II DNA/RNA helicases, SNF2 family [Transcription / DNA replication, recombination, and repair] $3x10^{-10}$ / COG5574, PEX10, RING-finger-containing E3 ubiquitin ligase [Posttranslational modification, protein turnover, chaperones] $4x10^{-5}$; PF00176, SNF2 family N-terminal domain, $2x10^{-14}$; PF00097, Zinc finger, C3HC4 type (RING finger), $1x10^{-14}$	2	early	-	32	R592, $2x10^{-14}$ R311, $3x10^{-8}$	ADO67436
crov403	DnaK/Hsp70	COG0443, DnaK, Molecular chaperone	0	early	-	74	L393, 0.0	ADO67437

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		[Posttranslational modification, protein turnover, chaperones] $7x10^{-164}$; PF00012, Hsp70 protein, 0.0					L254, $1x10^{-131}$	
crov404	unknown	-	6	late	+	-	-	ADO67438
crov405	unknown	-	6	-	-	-	-	ADO67439
crov406	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $3x10^{-13}$	-	-	-	6	L60, $2x10^{-39}$ and others	ADO67440
crov407	ribonuclease HII	COG0164, RnhB, Ribonuclease HII [DNA replication, recombination, and repair] $3x10^{-45}$; PF01351, Ribonuclease HII, $1x10^{-38}$	6	-	-	-	-	ADO67441
crov408	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $5x10^{-10}$	-	-	-	6	L414, $3x10^{-12}$ and others	ADO67442
crov409	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $2x10^{-15}$	48	-	-	6	L60, $2x10^{-52}$ and others	ADO67443
crov410	unknown	-	0	-	+	-	-	ADO67444
crov411	unknown	-	12	late	+	-	-	ADO67445
crov412	unknown	-	-	-	-	-	-	ADO67446
crov413	cysteine dioxygenase type I	PF05995, Cysteine dioxygenase type I, $1x10^{-6}$	unclear	early	-	-	-	ADO67447
crov414	unknown	-	6	-	-	-	-	ADO67448
crov415	unknown	-	6	-	-	-	-	ADO67449
crov416	unknown	-	N/A	early, late	-	-	-	ADO67450
crov417	contains SET domain	COG2940, COG2940, Proteins containing SET domain [General function prediction only] $5x10^{-4}$; PF00856, SET domain, $2x10^{-4}$	-	-	+	-	-	ADO67451

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov418	unknown	-	N/A	-	-	-	-	ADO67452
crov419	unknown	-	-	late	+	-	-	ADO67453
crov420	unknown	-	0	early	-	-	-	ADO67454
crov421	unknown	-	N/A	early	-	-	-	ADO67455
crov422	unknown	-	0	early	-	-	-	ADO67456
crov423	unknown	-	3	early	-	-	-	ADO67457
crov424	contains band 7 flotiilin domain	COG2268, COG2268, Uncharacterised protein conserved in bacteria [Function unknown] $7x10^{-11}$; PF01145, SPFH domain / Band 7 family, $9x10^{-8}$	2	early	-	-	-	ADO67458
crov425	unknown	PF09945, Predicted membrane protein (DUF2177), $8x10^{-10}$	-	-	+	1295	-	ADO67459
crov426	UvrD/REP helicase	COG0210, UvrD, Superfamily I DNA and RNA helicases [DNA replication, recombination, and repair] $1x10^{-68}$; PF00580, UvrD/REP helicase, $3x10^{-47}$	6	-	-	335	R568/R569, $3x10^{-10}$	ADO67460
crov427	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $2x10^{-13}$	2	early	-	6	L60, $3x10^{-43}$ and others	ADO67461
crov428	unknown	-	N/A	-	-	-	-	ADO67462
crov429	contains ankyrin repeats	COG0666, Arp, FOG: Ankyrin repeat [General function prediction only] $4x10^{-8}$; PF00023, Ankyrin repeat, $3x10^{-8}$	3	-	-	-	L93, $1x10^{-21}$ and others	ADO67463
crov430	contains ankyrin repeats	PF00023, Ankyrin repeat, $9x10^{-5}$	N/A	-	-	-	L289, $2x10^{-7}$ and others	ADO67464
crov431	unknown	PF03133, Tubulin-tyrosine ligase family, $2x10^{-5}$	-	-	-	-	-	ADO67465
crov432	unknown	-	3	-	+	-	-	ADO67466

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov433	unknown	-	0	-	+	-	-	ADO67467
crov434	unknown	-	N/A	-	+	-	-	ADO67468
crov435	ubiquitin-activating enzyme E1	COG0476, ThiF, Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2 [Coenzyme metabolism] $1x10^{-25}$, cd01490, Ube1_repeat2, Ubiquitin activating enzyme (E1) $5x10^{-104}$; PF00899, ThiF family, $5x10^{-33}$; PF02134, Repeat in ubiquitin-activating (UBA) protein, $8x10^{-12}$; PF09258, Ubiquitin-activating enzyme E1 C-terminal domain, $3x10^{-6}$	0	early	-	-	-	ADO67469
crov436	unknown	-	-	late	+	-	-	ADO67470
crov437	Rpb7, DNA-directed RNA polymerase II subunit 7	PF03876, RNA polymerase Rpb7-like; N-terminal domain, $6x10^{-7}$	2	early, late	+	1083	**L376, $3x10^{-19}$**	ADO67471
crov438	unknown	-	N/A	early	-	-	-	ADO67472
crov439	Rpb5, DNA-directed RNA polymerase II subunit 5	CT 3rd: COG2012, RPB5, DNA-directed RNA polymerase, subunit H, RpoH/RPB5 [Transcription] $1x10^{-7}$; PF01191, RNA polymerase Rpb5; C-terminal domain, $6x10^{-12}$	6	early	+	273	**L235, $3x10^{-13}$**	ADO67473
crov440	contains ankyrin repeats	PF00023, Ankyrin repeat, $3x10^{-4}$	-	-	-	-	L371, $7x10^{-76}$	ADO67474
crov441	unknown	PF05725, FNIP Repeat, $5x10^{-5}$	N/A	early	-	-	-	ADO67475
crov442	FtsJ-like methyltransferase	PF01728, FtsJ-like methyltransferase, $9x10^{-7}$	24	early	+	59	L511, $9x10^{-13}$ **R383, $4x10^{-12}$**	ADO67476
crov443	unknown	-	N/A	-	-	-	-	ADO67477
crov444	thiol oxidoreductase	COG5054, Erv1 / Alr family, $1x10^{-11}$; PF04777, Erv1 / Alr family, $5x10^{-18}$	N/A	-	+	52	R368, $5x10^{-19}$ **R596, $2x10^{-14}$**	ADO67478

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov445	unknown	-	-	-	-	-	-	ADO67479
crov446	NUDIX hydrolase	PF00293, NUDIX domain, $6x10^{-4}$	-	late	-	236	L375, $1x10^{-35}$	ADO67480
crov447	DNA methyltransferase	-	-	-	-	-	-	ADO67481
crov448	unknown	-	N/A	early	-	-	-	ADO67482
crov449	mitochondrial carrier protein	PF00153, Mitochondrial carrier protein, $3x10^{-10}$	12	-	-	-	-	ADO67483
crov450	unknown	-	6	-	-	-	-	ADO67484
crov451	unknown	-	6	late	+	-	-	ADO67485
crov452	ribonucleoside-dipohosphate reductase small subunit	COG0208, NrdF, Ribonucleotide reductase, beta subunit [Nucleotide transport and metabolism] $8x10^{-62}$; PF00268, Ribonucleotide reductase; small chain, $9x10^{-125}$	2	early	-	276	L312, $8x10^{-82}$	ADO67486
crov453	unknown	-	6	-	+	-	-	ADO67487
crov454	ribonucleoside-diphosphate reductase LSU, intein-containing	COG0209 NrdA, Ribonucleotide reductase, alpha subunit [Nucleotide transport and metabolism] $3x10^{-46}$; PF03477, ATP cone domain, $9x10^{-10}$; PF00317, Ribonucleotide reductase, all-alpha domain, $7x10^{-27}$; PF02867, Ribonucleotide reductase, barrel domain, $1x10^{-168}$; PF00462, Glutaredoxin, $1x10^{-10}$	0	early	-	1353	R313, $5x10^{-71}$	ADO67488
crov455	unknown	-	0	early	-	-	-	ADO67489
crov456	unknown	-	-	-	+	-	L371, $8x10^{-14}$	ADO67490
crov457	ubiquitin carboxyl-terminal hydrolase	COG5533, UBP5, Ubiquitin C-terminal hydrolase [Posttranslational modification, protein turnover, chaperones] $4x10^{-33}$; PF00443, Ubiquitin carboxyl-terminal	3	early	-	331	R319, $4x10^{-29}$	ADO67491

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		hydrolase, $1x10^{-51}$						
crov458	family X DNA polymerase	COG1796, POL4, DNA polymerase IV (family X) [DNA replication, recombination, and repair] $3x10^{-29}$; PF10391, Fingers domain of DNA polymerase lambda, $4x10^{-8}$	3	early	+	4	**L318, $2x10^{-58}$**	ADO67492
crov459	metal-dependent hydrolase	COG1234, ElaC, Metal-dependent hydrolases of the beta-lactamase superfamily III [General function prediction only] $5x10^{-12}$;	6	early	-	-	-	ADO67493
crov460	unknown	-	-	late	-	-	-	ADO67494
crov461	replication factor C small subunit	COG2812, DnaX, DNA polymerase III, gamma/tau subunits [DNA replication, recombination, and repair] $1x10^{-13}$; PF00004, ATPase family associated with various cellular activities (AAA), $7x10^{-5}$	6	early, late	-	1351	R395, $3x10^{-25}$ L510, $3x10^{-24}$ L499, $2x10^{-22}$	ADO67495
crov462	NAD-dependent DNA ligase	COG0272, Lig, NAD-dependent DNA ligase (contains BRCT domain type II) [DNA replication, recombination, and repair] $5x10^{-68}$; PF01653; PF03120; PF00533, NAD-dependent DNA ligase adenylation domain; NAD-dependent DNA ligase OB-fold domain; BRCA1 C Terminus (BRCT) domain, $5x10^{-11}$; $2x10^{-16}$; $2x10^{-10}$	2	-	+	35	R303, $3x10^{-112}$	ADO67496
crov463	unknown	-	N/A	-	-	-	-	ADO67497
crov464	unknown	-	N/A	-	-	-	-	ADO67498
crov465	unknown	-	-	late	+	-	-	ADO67499
crov466	unknown	-	-	-	+	-	-	ADO67500
crov467	tetratrico peptide repeats	PF08238, Sel1 repeat, 0.03	N/A	-	-	-	-	ADO67501

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov468	tetratrico peptide repeats	COG0790, FOG: TPR repeat, SEL1 subfamily [General function prediction only] $2x10^{-6}$; PF08238, Sel1 repeat, $7x10^{-6}$	6	-	-	-	L21, $3x10^{-6}$	ADO67502
crov469	tetratrico peptide repeats	PF07720, Tetratricopeptide repeat, 0.04	N/A	early	-	-	-	ADO67503
crov470	unknown	-	N/A	early	-	-	-	ADO67504
crov471	tetratrico peptide repeats	COG0790, FOG: TPR repeat, SEL1 subfamily [General function prediction only] $3x10^{-8}$; PF08238, Sel1 repeat, $3x10^{-6}$	N/A	early	-	-	-	ADO67505
crov472	contains methytransferase domain	PF07669, Eco57I restriction endonuclease, 0.003	3	early	-	-	-	ADO67506
crov473	unknown	-	N/A	-	-	-	-	ADO67507
crov474	unknown	-	6	early	-	558	L308, $3x10^{-9}$	ADO67508
crov475	family 2C Ser/Thr phosphatase	COG0631, PTC1, Serine/threonine protein phosphatase [Signal transduction mechanisms] $2x10^{-26}$; PF00481, Protein phosphatase 2C, $9x10^{-54}$	6	-	+	1061	**R306/R307, $8x10^{-40}$**	ADO67509
crov476	unknown	-	2	early, late	-	-	-	ADO67510
crov477	unknown	-	6	-	-	-	-	ADO67511
crov478	Lon protease	COG0466, Lon, ATP-dependent Lon protease, bacterial type [Posttranslational modification, protein turnover, chaperones] $5x10^{-86}$; PF00004, ATPase family associated with various cellular activities (AAA), $8x10^{-15}$; PF05362, Lon protease (S16) C-terminal proteolytic domain, $5x10^{-31}$	6	-	-	228	L251, $6x10^{-102}$	ADO67512

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov479	eIF-2gamma	COG5257 GCD11, Translation initiation factor 2, gamma subunit (eIF-2gamma; GTPase) [Translation, ribosomal structure and biogenesis] $3x10^{-69}$; PF00009, Elongation factor Tu GTP binding domain, $5x10^{-30}$	2	early	-	-	-	ADO67513
crov480	unknown	-	0	-	-	1088	-	ADO67514
crov481	unknown	-	0, cross-hybridization with host mRNA	-	-	-	-	ADO67515
crov482	Rpb3/Rpb11, DNA-directed RNA polymerase II subunit 3/11	CT 3rd: COG1761, RPB11, DNA-directed RNA polymerase, subunit L [Transcription] $5x10^{-6}$; PF01193, RNA polymerase Rpb3/Rpb11 dimerisation domain, $3x10^{-4}$	0	early	+	635	**R470, $1x10^{-10}$**	ADO67516
crov483	unknown	-	6	-	-	-	-	ADO67517
crov484	ribonuclease H1	COG0328, RnhA, Ribonuclease HI [DNA replication, recombination, and repair] $5x10^{-16}$; PF00075, RNase H, $4x10^{-24}$	-	late	-	1352	R298/R299, $8x10^{-21}$	ADO67518
crov485	cysteine protease	PF02338, OTU-like cysteine protease, $5x10^{-10}$	-	-	+	-	-	ADO67519
crov486	DNA mismatch repair protein MutS	COG0249 MutS, Mismatch repair ATPase (MutS family) [DNA replication, recombination, and repair] $2x10^{-79}$; PF01624, MutS domain I, $2x10^{-7}$; PF05192, MutS domain III, $3x10^{-7}$; PF00488, MutS domain V, $3x10^{-32}$	6	-	-	72	L359, $6x10^{-116}$	ADO67520
crov487	unknown	-	-	-	+	-	-	ADO67521
crov488	unknown	-	-	-	+	-	-	ADO67522
crov489	Ser/Thr protein kinase	COG0515, SPS1, Serine/threonine protein kinase [General function prediction only / Signal transduction mechanisms /	2	early	-	304	L205, $3x10^{-12}$	ADO67523

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
		Transcription / DNA replication, recombination, and repair] $3x10^{-6}$; PF00069, Protein kinase domain, $2x10^{-19}$						
crov490	unknown	-	N/A	early	-	-	-	ADO67524
crov491	Rpb6, DNA-directed RNA polymerase II subunit 6	COG1758, RpoZ, DNA-directed RNA polymerase, subunit K/omega [Transcription] 0.004; PF01192, RNA polymerase Rpb6, $4x10^{-7}$	2	early	+	522	R209, $3x10^{-12}$	ADO67525
crov492	Rpb9, DNA-directed RNA polymerase II subunit 9	-	1	early	+	521	**L208, $6x10^{-13}$**	ADO67526
crov493	ubiquitin-conjugating enzyme E2	COG5078, Ubiquitin-protein ligase [Posttranslational modification, protein turnover, chaperones] $2x10^{-12}$; PF00179, Ubiquitin-conjugating enzyme, $6x10^{-32}$	-	-	-	329	R521, $5x10^{-21}$ R795, $2x10^{-9}$ L709, $1x10^{-6}$	ADO67527
crov494	VV D5-like primase/helicase	COG3378, COG3378, Predicted ATPase [General function prediction only] $1x10^{-30}$; PF08707, Primase C terminal 2 (PriCT-2), $4x10^{-21}$; PF08706, D5 N terminal-like, $4x10^{-13}$	2	early	-	23	L206/L207, $2x10^{-118}$	ADO67528
crov495	unknown	-	-	-	-	986	R217, $1x10^{-33}$	ADO67529
crov496	unknown	-	-	late	+	-	-	ADO67530
crov497	B family DNA polymerase, intein-containing	COG0417 PolB, DNA polymerase elongation subunit (family B) [DNA replication, recombination, and repair] $3x10^{-36}$; PF03104;, DNA polymerase family B, exonuclease domain, $1x10^{-48}$; PF00136, DNA polymerase family B, $4x10^{-48}$	3	early	-	38	R322, $3x10^{-101}$	ADO67531
crov498	unknown	-	0	-	+	-	-	ADO67532
crov499	unknown	-	N/A	early	-	-	-	ADO67533

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov500	carrier protein	PF00153, Mitochondrial carrier protein, $3x10^{-4}$	0	early	-	-	L276, $1x10^{-8}$	ADO67534
crov501	unknown	-	6	-	-	-	-	ADO67535
crov502	bifunctional dihydrofolate reductase/thymidylate synthase	COG0262, FolA, Dihydrofolate reductase [Coenzyme metabolism] $2x10^{-19}$ / COG0207, ThyA, Thymidylate synthase [Nucleotide transport and metabolism] $3x10^{-56}$; PF00186, Dihydrofolate reductase, $2x10^{-33}$; PF00303, Thymidylate synthase, $2x10^{-106}$	3	early	-	1136	R497, $5x10^{-55}$	ADO67536
crov503	helicase/exonuclease	COG0210, UvrD, Superfamily I DNA and RNA helicases [DNA replication, recombination, and repair] $1x10^{-5}$ / COG0847, DnaQ, DNA polymerase III, epsilon subunit and related 3'-5' exonucleases [DNA replication, recombination, and repair] $2x10^{-5}$; PF00929, Exonuclease, $9x10^{-15}$	-	-	-	47	R568/R569, $6x10^{-8}$	ADO67537
crov504	unknown	-	-	-	-	-	-	ADO67538
crov505	isoleucyl-tRNA synthetase	COG0060 IleS, Isoleucyl-tRNA synthetase [Translation, ribosomal structure and biogenesis] $2x10^{-163}$; PF00133, tRNA synthetases class I (I, L, M and V), $5x10^{-148}$; PF08264, Anticodon-binding domain, $1x10^{-22}$	0	early	-	-	-	ADO67539
crov506	SFII helicase	COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $2x10^{-7}$; PF00097, Zinc finger, C3HC4 type (RING finger), $3x10^{-8}$	1	early	-	32	R592, $1x10^{-6}$	ADO67540
crov507	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $4x10^{-13}$	N/A	early	-	6	L60, $4x10^{-27}$ and others	ADO67541
crov508	unknown	-	0	-	-	-	-	ADO67542
crov509	unknown	-	2	early	-	-	-	ADO67543

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov510	unknown	-	N/A	-	-	-	-	ADO67544
crov511	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $2x10^{-13}$	0	-	-	6	L60, $3x10^{-46}$ and others	ADO67545
crov512	unknown	-	6	early	-	-	-	ADO67546
crov513	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $3x10^{-8}$	N/A	-	-	6	L281, $2x10^{-7}$ and others	ADO67547
crov514	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $4x10^{-13}$	-	early	-	6	L60, $1x10^{-37}$ and others	ADO67548
crov515	type I restriction modification enzyme, M and S domains	COG0286, HsdM, Type I restriction-modification system methyltransferase subunit [Defence mechanisms] $4x10^{-5}$ / COG0732, HsdS, Restriction endonuclease S subunits [Defence mechanisms] $9x10^{-4}$; PF02384, N-6 DNA Methylase, $2x10^{-4}$; PF01420, Type I restriction modification DNA specificity domain, $2x10^{-10}$	0	early	-	-	-	ADO67549
crov516	unknown	-	-	-	-	-	-	ADO67550
crov517	unknown	-	N/A	-	-	-	-	ADO67551
crov518	unknown	-	N/A	-	-	-	-	ADO67552
crov519	unknown	-	-	-	-	-	-	ADO67553
crov520	unknown	-	-	-	-	-	-	ADO67554
crov521	unknown	-	N/A	early	-	-	-	ADO67555
crov522	unknown	PF01541, GIY-YIG catalytic domain, $2x10^{-4}$	N/A	-	-	-	-	ADO67556
crov523	unknown	-	-	-	-	-	-	ADO67557
crov524	unknown	-	N/A	-	-	-	-	ADO67558
crov525	Ser/Thr protein	PF00149, Calcineurin-like phosphoesterase,	0	-	+	-	-	ADO67559

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
	phosphatase	$5x10^{-10}$						
crov526	SFII helicase/restriction enzyme	COG1061, SSL2, DNA or RNA helicases of superfamily II [Transcription / DNA replication, recombination, and repair] $2x10^{-12}$; PF04851, Type III restriction enzyme, res subunit, $7x10^{-8}$	12	early	-	-	-	ADO67560
crov527	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $1x10^{-13}$	N/A	early	-	6	L60, $4x10^{-58}$ and others	ADO67561
crov528	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $7x10^{-14}$	1	early	-	6	L60, $2x10^{-42}$ and others	ADO67562
crov529	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $3x10^{-13}$	N/A	early	-	6	R636, $2x10^{-10}$ and others	ADO67563
crov530	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $5x10^{-14}$	N/A	-	-	6	L60, $1x10^{-36}$ and others	ADO67564
crov531	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $2x10^{-10}$	N/A	early	-	6	R636, $4x10^{-10}$ and others	ADO67565
crov532	unknown	-	6	early	-	-	-	ADO67566
crov533	unknown	-	3	-	-	-	-	ADO67567
crov534	unknown	-	N/A	-	-	-	-	ADO67568
crov535	unknown	-	-	-	-	-	-	ADO67569
crov536	unknown	-	N/A	early	-	-	-	ADO67570
crov537	contains peptidase S41 domain	PF03572, Peptidase family S41, $1x10^{-5}$	-	-	-	-	-	ADO67571
crov538	FNIP/IP22-like repeats	-	N/A	-	-	-	-	ADO67572
crov539	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $9x10^{-14}$	N/A	early	-	6	L60, $7x10^{-27}$ and others	ADO67573

CDS	Predicted function or protein domain	COG or Pfam Database Matches (Domain ID, Description, E-value)	Microarray (h p.i.)	Promoter type	Virion protein	NCVOG	Mimivirus (Gene ID, E-value)	GenBank Acc. No.
crov540	unknown	-	N/A	early	-	-	-	ADO67574
crov541	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $3x10^{-16}$	N/A	early	-	6	L60, $3x10^{-28}$ and others	ADO67575
crov542	FNIP/IP22-like repeats	PF05725, FNIP Repeat, $8x10^{-7}$	2	early	-	6	L60, $1x10^{-11}$ and others	ADO67576
crov543	unknown	-	-	-	-	-	-	ADO67577

Appendix C: Physical properties of CroV CDSs

Table C.1 Physical properties of CroV CDSs

Location refers to the nucleotide coordinates of the CDSs in the CroV genome sequence as deposited in GenBank under the accession number GU244497. MW lists the predicted molecular weight in kilodaltons. Transmembrane helices were predicted by TMHMM server 2.0[17] and signal peptides and anchors by SignalP[18].

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov001	458-42	5	35.01	138	15.43	4.48	-	-
crov002	2298-514	6	22.46	572	69.43	4.89	-	-
crov003	4371-4538	3	41.07	55	6.42	9.28	-	-
crov004	4741-5157	1	18.23	138	16.25	8.81	-	-
crov005	6041-5814	5	21.05	75	9.09	11.13	-	-
crov006	6177-6584	3	19.36	135	15.77	10.25	-	-
crov007	6646-7029	1	20.31	127	15.19	9.8	-	-
crov008	7436-7026	5	22.87	136	16.11	5.26	-	-
crov009	7913-7482	5	20.14	143	17.02	6.26	-	-
crov010	8840-8049	5	17.93	263	31.72	9.37	-	-
crov011	9796-8933	4	22.45	287	34.53	9.73	-	-
crov012	11047-9843	4	33.61	401	44.99	4.92	-	-
crov013	11114-12553	2	20.42	479	55.96	9.74	1	-
crov014	12570-12950	3	19.69	126	14.28	9.97	3	peptide
crov015	13039-15780	1	23.56	913	106.32	8.89	-	-
crov016	15889-16056	1	18.45	55	6.36	10.32	-	-
crov017	16117-16356	1	14.58	79	9.40	10.08	-	-
crov018	18410-16353	5	19.68	685	82.32	8.77	-	-
crov019	18896-20368	2	24.85	490	58.44	7.82	-	-
crov020	20979-20371	6	16.58	202	24.74	8.03	-	-
crov021	22192-21203	4	18.79	329	39.58	8.58	-	-
crov022	22787-22467	5	20.25	106	13.04	9.69	-	-
crov023	22977-23537	3	19.07	186	22.51	9.61	-	-
crov024	23728-23967	1	18.33	79	9.40	10.11	-	-
crov025	24223-24582	1	12.50	119	14.71	10.19	-	-

[17] http://www.cbs.dtu.dk/services/TMHMM
[18] http://www.cbs.dtu.dk/services/SignalP

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov026	24599-25483	2	18.42	294	34.99	9.69	-	-
crov027	26847-25852	6	20.28	331	38.98	6.37	-	-
crov028	27514-27281	4	20.09	77	9.29	4.36	-	-
crov029	28828-28274	4	19.10	184	21.68	10.15	-	-
crov030	29219-29034	5	9.68	61	7.64	10.37	-	-
crov031	30962-29727	5	22.65	411	47.40	5.25	-	-
crov032	31485-31003	6	20.29	160	18.95	9.77	-	-
crov033	31842-31609	6	17.95	77	9.53	5.88	-	-
crov034	32391-31876	6	27.52	171	19.37	11.02	-	-
crov035	34878-32416	6	17.50	820	98.33	8.58	-	-
crov036	37306-34871	4	26.68	811	93.07	7.34	-	-
crov037	37404-38369	3	18.63	321	37.62	9.61	-	-
crov038	40478-38364	5	17.68	704	84.67	9.69	-	-
crov039	40520-40981	2	30.95	153	17.75	9.87	-	-
crov040	41055-41291	3	19.83	78	8.92	8.93	2	peptide
crov041	42268-41288	4	20.80	326	38.78	7.42	-	-
crov042	42532-42341	4	17.71	63	7.87	4.84	1	-
crov043	44373-42616	6	24.06	585	67.34	10.19	-	-
crov044	44659-44441	4	33.33	72	8.49	5.53	-	-
crov045	45073-44726	4	19.54	115	13.32	10.71	-	-
crov046	45680-45450	5	22.08	76	8.98	10.34	-	-
crov047	45851-48742	2	28.67	963	106.45	4.21	-	-
crov048	48788-50137	2	19.70	449	52.66	5.47	-	-
crov049	50505-50134	6	15.32	123	15.31	8.33	-	-
crov050	50796-50536	6	17.24	86	10.62	4.85	-	-
crov051	51852-50836	6	19.76	338	39.54	10.05	-	-
crov052	51929-52279	2	15.67	116	14.00	9.89	-	-
crov053	52327-52737	1	21.65	136	16.06	9.53	-	-
crov054	52766-53326	2	21.39	186	21.26	5.09	-	-
crov055	53345-53788	2	20.50	147	17.18	7.89	-	-
crov056	54382-53789	4	21.89	197	22.99	10.05	-	-
crov057	56362-54416	4	14.18	648	78.41	9.82	-	-
crov058	56875-56399	4	27.04	158	18.27	6.25	-	-
crov059	58185-56971	6	50.95	404	43.16	6.3	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov060	58997-58275	5	20.06	240	28.49	7.7	1	peptide
crov061	60159-59002	6	23.06	385	45.63	9.55	2	-
crov062	61272-60184	6	16.90	362	43.30	9.48	-	-
crov063	61634-61308	5	22.63	108	12.59	6.52	-	-
crov064	62220-61681	6	21.48	179	21.55	9.24	-	-
crov065	62371-62919	1	19.31	182	21.56	8.44	-	-
crov066	63290-63018	5	23.81	90	10.27	6.22	-	-
crov067	65315-63450	5	24.97	621	72.68	6.1	-	-
crov068	67248-65365	6	24.47	627	73.49	5.68	-	-
crov069	67775-67284	5	30.49	163	17.84	8.3	-	-
crov070	68224-67865	4	20.56	119	13.99	8.93	-	-
crov071	69566-68835	5	22.95	243	29.15	9.19	-	-
crov072	69645-71966	3	22.83	773	90.81	8.48	-	-
crov073	72036-72686	3	17.82	216	25.09	8.8	-	-
crov074	73560-72667	6	21.48	297	34.61	10.16	-	-
crov075	73658-75328	2	26.09	556	65.07	7.96	-	-
crov076	75747-75340	6	16.42	135	16.30	10.14	-	-
crov077	75797-76066	2	22.59	89	10.34	8.49	-	-
crov078	76112-76483	2	25.00	123	14.93	9.38	-	-
crov079	76486-77931	1	24.48	481	54.67	8.97	2	-
crov080	78526-77963	4	24.47	187	21.85	9.09	-	-
crov081	78585-78803	3	27.40	72	8.05	6.49	2	peptide
crov082	79314-78820	6	28.89	164	18.89	9.38	-	-
crov083	80090-79362	5	25.65	242	28.05	6.92	8	-
crov084	82350-80107	6	23.71	747	86.35	10.02	-	-
crov085	82427-82717	2	23.37	96	11.20	8.03	1	peptide
crov086	82919-82725	5	17.44	64	7.80	7.68	-	-
crov087	83567-82944	5	23.72	207	24.32	7.4	-	-
crov088	85772-83610	5	27.37	720	80.46	4.75	1	-
crov089	85809-87155	3	24.28	448	49.32	4.67	1	-
crov090	87185-88471	2	28.28	428	45.66	4.16	1	-
crov091	89235-88468	6	23.18	255	29.93	6.74	7	-
crov092	89322-89819	3	22.09	165	19.55	7.32	-	-
crov093	90373-89903	4	21.02	156	18.80	6.88	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov094	90494-91081	2	30.27	195	21.80	11.09	-	-
crov095	92060-91083	5	28.02	325	38.42	9.7	-	-
crov096	92652-92107	6	32.97	181	21.60	9.68	-	-
crov097	92713-93495	1	20.95	260	30.59	9.83	-	-
crov098	93723-93484	6	24.58	79	9.04	10.37	-	-
crov099	93785-94354	2	29.65	189	22.29	6	-	-
crov100	95461-94427	4	22.03	344	40.55	9.7	-	-
crov101	95542-95910	1	26.83	122	14.20	7.51	-	-
crov102	96429-95977	6	31.57	150	16.80	8.16	-	-
crov103	97093-96473	4	21.26	206	23.78	9.74	-	-
crov104	97195-97428	1	29.06	77	9.31	4.25	-	-
crov105	97468-98046	1	23.14	192	22.52	3.67	3	-
crov106	98853-98065	6	25.73	262	30.55	9.74	-	-
crov107	100308-98899	6	21.35	469	55.26	8.8	-	-
crov108	100650-100438	6	23.94	70	8.34	9.2	-	-
crov109	104103-100708	6	25.80	1131	132.21	9.2	-	-
crov110	105137-104148	5	27.07	329	36.21	6.67	-	-
crov111	107482-105167	4	25.60	771	89.19	8.59	-	-
crov112	108118-107534	4	19.83	194	22.80	9.99	-	-
crov113	108207-110051	3	25.69	614	68.82	9.29	-	-
crov114	112168-110390	4	24.90	592	69.13	6.1	-	-
crov115	113787-112363	6	21.68	474	56.68	10.15	-	-
crov116	113878-114234	1	21.85	118	13.88	5.15	-	-
crov117	115142-114195	5	20.57	315	36.25	9.45	-	-
crov118	120111-115186	6	21.32	1641	193.75	9.39	-	-
crov119	120803-120144	5	16.82	219	27.08	10.12	-	-
crov120	120834-121619	3	19.08	261	31.28	9.72	5	-
crov121	121714-122376	1	22.47	220	25.85	9.11	-	-
crov122	123255-123091	6	20.61	54	6.24	9	-	-
crov123	123749-123348	5	20.40	133	15.39	7.51	-	-
crov124	124589-123777	5	21.28	270	32.03	8.93	-	-
crov125	125236-124607	4	17.62	209	25.03	10.59	-	-
crov126	126300-125269	6	28.78	343	40.55	9.27	-	-
crov127	127586-126378	5	31.02	402	44.77	9.2	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov128	127683-128786	3	24.64	367	43.35	8.5	-	-
crov129	128852-129268	2	17.51	138	16.51	8.96	-	-
crov130	131169-129283	6	18.34	628	74.64	9.74	-	-
crov131	131226-132938	3	25.28	570	66.25	10.02	-	-
crov132	132981-134465	3	22.90	494	58.07	9.46	-	-
crov133	134499-135959	3	26.35	486	55.72	8.32	-	peptide
crov134	135981-136322	3	31.29	113	12.46	4.4	-	-
crov135	136425-137330	3	21.19	301	35.68	9.63	1	-
crov136	137466-139163	3	28.56	565	63.94	9.4	-	-
crov137	141763-139160	4	18.93	858	102.12	9.52	-	-
crov138	141774-142565	3	18.43	263	30.80	10.08	-	-
crov139	144740-142557	5	17.58	727	87.41	6.46	-	-
crov140	146541-144757	6	19.83	594	70.27	9.73	-	-
crov141	146628-147110	3	25.67	160	18.78	4.87	1	-
crov142	147186-152738	3	24.94	1850	210.28	4.37	-	-
crov143	153288-152743	6	26.37	181	21.36	9.27	-	-
crov144	153894-153337	6	21.15	185	22.42	8.72	-	-
crov145	153979-154746	1	27.60	255	29.24	10.06	-	-
crov146	154804-155580	1	22.78	258	30.45	8.92	-	-
crov147	156641-155577	5	22.91	354	40.68	6.14	-	-
crov148	156700-157428	1	32.37	242	26.22	6.23	-	-
crov149	157460-158902	2	22.45	480	58.23	10.1	-	-
crov150	161339-158907	5	23.22	810	94.17	8.99	-	-
crov151	161592-162872	3	27.01	426	48.38	8.43	-	-
crov152	162913-164517	1	24.17	534	62.14	10.29	-	-
crov153	167636-164568	5	38.16	1022	117.27	8.68	-	-
crov154	167751-168062	3	24.04	103	12.17	10.86	-	-
crov155	170443-168155	4	38.93	762	85.50	7.03	-	-
crov156	171214-170528	4	36.24	228	25.36	4.7	-	-
crov157	172652-171366	5	43.90	428	46.47	5.63	-	-
crov158	172918-172688	4	35.93	76	8.03	9.04	1	peptide
crov159	174420-172999	6	34.46	473	52.94	4.82	1	-
crov160	175418-174519	5	22.89	299	34.40	10.55	-	-
crov161	176124-175558	6	22.05	188	21.86	8.75	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov162	176956-176171	4	20.61	261	30.99	8.82	-	-
crov163	177051-177908	3	20.98	285	33.69	9.54	-	-
crov164	178619-177900	5	23.06	239	27.89	9.71	-	-
crov165	178693-179259	1	17.64	188	21.79	4.92	-	-
crov166	180563-179301	5	28.82	420	47.97	9.28	-	-
crov167	180638-180997	2	17.50	119	14.09	8.89	1	-
crov168	181281-180994	6	21.18	95	11.40	8.63	-	-
crov169	181315-182172	1	20.16	285	33.01	8.41	2	-
crov170	182929-182174	4	24.34	251	29.68	7.62	-	-
crov171	183080-184114	2	25.89	344	39.91	6.51	-	-
crov172	184570-184109	4	22.73	153	17.49	6.5	1	peptide
crov173	184609-185412	1	15.17	267	32.14	9.53	-	-
crov174	185435-186700	2	27.73	421	46.85	6.95	-	-
crov175	186761-187525	2	29.28	254	28.07	9.63	1	-
crov176	189047-187527	5	21.76	506	58.77	7.92	-	-
crov177	189667-189164	4	33.33	167	19.36	7.26	-	-
crov178	199740-189727	6	18.89	3337	394.97	5.04	-	-
crov179	199800-200141	3	18.42	113	13.18	10.25	-	-
crov180	200168-200425	2	22.09	85	9.68	4.47	-	-
crov181	200489-200824	2	20.83	111	13.15	9.53	-	-
crov182	200859-201947	3	22.59	362	42.66	6.21	-	-
crov183	201994-204324	1	25.95	776	90.30	6.26	-	-
crov184	205198-204314	4	23.73	294	34.71	8.67	-	-
crov185	207877-205268	4	29.73	869	98.74	9.94	-	-
crov186	208530-207946	6	32.65	194	22.12	7.52	-	-
crov187	209024-208644	5	28.35	126	14.42	8.84	-	-
crov188	209952-209047	6	16.34	301	36.32	8.9	-	-
crov189	210049-210939	1	24.24	296	35.01	9.68	-	-
crov190	211007-212254	2	21.15	415	49.17	7.88	-	-
crov191	212311-212778	1	18.16	155	18.70	9.59	-	-
crov192	213427-212915	4	23.00	170	20.20	9.79	-	-
crov193	213535-214485	1	24.82	316	36.97	8.99	-	-
crov194	215076-214507	6	23.33	189	22.14	7.49	2	peptide
crov195	215151-216095	3	20.00	314	37.08	9.47	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov196	216137-217099	2	20.25	320	37.05	9.65	-	-
crov197	217139-217405	2	19.10	88	11.01	4.92	-	-
crov198	217440-219629	3	23.15	729	84.69	8.82	-	-
crov199	219647-221776	2	28.92	709	81.58	7.5	-	peptide
crov200	221823-222893	3	19.42	356	41.72	9.79	-	-
crov201	223119-222886	6	29.06	77	9.07	7.8	-	-
crov202	223190-223501	2	19.87	103	12.16	5.21	1	-
crov203	225327-223507	6	35.97	606	65.36	3.4	-	-
crov204	225234-225656	3	17.02	140	16.88	9.88	2	peptide
crov205	225681-226088	3	17.65	135	15.91	9.11	-	-
crov206	226141-227580	1	25.76	479	57.36	6.63	-	-
crov207	227936-227577	5	23.89	119	13.79	8.72	-	-
crov208	228014-228610	2	25.63	198	22.56	8.43	-	-
crov209	228682-229290	1	27.59	202	23.64	9.75	-	-
crov210	229671-229303	6	18.43	122	14.71	9.46	-	-
crov211	229906-229712	4	18.46	64	7.46	9.73	-	-
crov212	229971-233018	3	24.74	1015	119.40	8.9	-	-
crov213	233074-233493	1	16.43	139	16.81	10.69	-	-
crov214	233537-234991	2	21.03	484	57.48	8.87	-	-
crov215	235530-234988	6	29.47	180	20.73	6.68	-	-
crov216	237471-235573	6	21.12	632	74.23	8.82	-	-
crov217	240099-237496	6	23.92	867	101.45	9.39	-	-
crov218	240177-241130	3	25.89	317	35.38	9.27	-	-
crov219	241184-242020	2	23.78	278	31.74	4.77	-	-
crov220	242514-242023	6	23.78	163	19.17	8.21	-	-
crov221	242586-243041	3	15.79	151	17.91	9.32	-	-
crov222	243061-244227	1	17.40	388	45.85	8.17	1	peptide
crov223	244266-245138	3	27.61	290	33.40	4.54	2	peptide
crov224	245212-249207	1	29.28	1331	152.40	8.53	-	-
crov225	250173-249277	6	20.51	298	35.01	8.86	-	-
crov226	250228-250656	1	26.81	142	16.32	10.07	-	-
crov227	251198-250653	5	28.21	181	20.27	3.65	1	-
crov228	251911-251267	4	30.85	214	23.76	3.71	1	peptide
crov229	251990-252361	2	24.19	123	13.98	8.84	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov230	252407-253543	2	24.89	378	43.21	9.25	-	-
crov231	254390-253548	5	22.89	280	32.72	9.26	-	-
crov232	254497-256041	1	35.66	514	56.94	8.1	3	-
crov233	256096-256860	1	23.01	254	30.28	5.67	-	-
crov234	257040-256870	6	22.81	56	6.64	4.69	1	peptide
crov235	257308-257102	4	23.19	68	8.30	8.9	1	peptide
crov236	257675-257355	5	29.91	106	11.56	3.7	1	peptide
crov237	260700-257893	6	22.93	935	107.51	8.07	-	-
crov238	260795-261484	2	36.81	229	25.62	10.58	-	-
crov239	261540-262406	3	30.33	288	32.48	4.48	-	-
crov240	262711-262475	4	35.86	78	8.37	5.54	1	peptide
crov241	264726-262741	6	21.20	661	76.84	4.46	-	-
crov242	264817-265650	1	24.82	277	31.95	7.55	-	-
crov243	265711-266715	1	19.50	334	39.64	9.58	-	-
crov244	267407-266730	5	26.55	225	25.62	5.61	-	-
crov245	267571-267413	4	23.90	52	6.14	3.56	-	-
crov246	268292-267648	5	15.66	214	26.11	9.96	-	-
crov247	268337-269032	2	20.11	231	27.14	9.73	-	-
crov248	269070-269243	3	16.09	57	7.01	9.27	-	-
crov249	269494-269270	4	20.44	74	8.59	9.86	-	-
crov250	269807-270727	2	21.72	306	36.38	9.83	-	-
crov251	270953-270693	5	17.62	86	10.38	9.06	-	-
crov252	271677-270964	6	16.11	237	28.99	9.6	-	-
crov253	271754-272695	2	18.47	313	37.22	10.02	-	-
crov254	273476-272715	5	22.31	253	30.44	9.35	-	-
crov255	274273-273512	4	19.16	253	30.15	9.71	-	-
crov256	276187-274283	4	19.32	634	75.09	9.57	-	-
crov257	276238-276396	1	20.13	52	6.17	9.57	-	-
crov258	276428-276799	2	22.85	123	14.02	7.42	-	-
crov259	277826-276852	5	18.36	324	38.76	8.91	-	-
crov260	278738-277836	5	20.27	300	35.55	9.42	-	-
crov261	278830-280422	1	23.48	530	61.99	8.78	-	-
crov262	280499-281314	2	18.87	271	32.21	9.8	-	-
crov263	281324-282067	2	20.83	247	29.69	9.05	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov264	282418-282768	1	17.09	116	13.56	10.03	-	-
crov265	282815-284188	2	18.20	457	52.76	8.37	-	-
crov266	285489-284185	6	20.00	434	50.14	7.35	-	-
crov267	286987-285527	4	25.53	486	56.35	7.74	-	-
crov268	287071-288369	1	22.25	432	50.64	9.11	-	-
crov269	289118-288360	5	20.42	252	29.71	6.66	-	-
crov270	291467-289122	5	14.75	781	93.73	9.51	-	-
crov271	293950-291485	4	16.63	821	97.61	9.59	-	-
crov272	294811-293993	4	19.54	272	32.96	9.5	-	-
crov273	300020-294858	5	20.61	1720	203.35	9	-	-
crov274	301387-300224	4	14.95	387	46.17	10.21	-	-
crov275	302466-301420	6	20.53	348	41.02	8.95	-	-
crov276	302560-303219	1	29.85	219	24.74	3.25	-	-
crov277	303260-303502	2	28.81	80	9.24	8.5	-	-
crov278	304354-303518	4	20.79	278	32.73	8.28	-	-
crov279	304430-305701	2	22.64	423	48.84	9.62	-	-
crov280	305807-306595	2	36.38	262	27.45	9.16	-	-
crov281	306638-307090	2	34.22	150	17.08	8.4	-	-
crov282	307821-307108	6	19.75	237	27.96	9.43	-	-
crov283	310191-307858	6	25.75	777	90.87	9.3	-	-
crov284	310905-310339	6	23.63	188	21.63	5.71	2	anchor
crov285	311851-311099	4	39.44	250	26.50	4.55	-	-
crov286	312157-313518	1	24.08	453	53.72	8.98	-	-
crov287	313600-313899	1	27.00	99	11.83	4.54	-	-
crov288	313938-314498	3	21.75	186	21.82	8.43	-	-
crov289	314591-314776	2	19.89	61	7.19	8.49	-	peptide
crov290	315781-314810	4	29.12	323	35.37	7.18	-	-
crov291	315841-316428	1	21.60	195	23.03	5.94	1	anchor
crov292	316460-321946	2	20.74	1828	217.13	9.45	-	-
crov293	322209-321985	6	24.44	74	8.69	6.81	-	-
crov294	322529-322224	5	17.97	101	11.68	5.75	2	-
crov295	322571-323086	2	21.90	171	20.33	6.51	-	-
crov296	323112-323708	3	24.96	198	23.19	4.69	-	-
crov297	323763-325829	3	19.74	688	81.21	7.82	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov298	325880-328516	2	22.53	878	103.83	8.98	-	-
crov299	328565-329038	2	25.11	157	18.48	9.55	-	-
crov299a	329188-329045	5	22.22	47	5.56	8.18	1	peptide
crov300	329652-329407	6	21.95	81	9.23	8.46	-	-
crov301	330090-329695	6	24.24	131	14.86	10.13	-	-
crov302	330177-331664	3	23.39	495	58.23	9.45	1	-
crov303	331709-332548	2	21.55	279	33.30	9.94	-	-
crov304	332653-334341	1	24.69	562	66.40	8.92	-	-
crov305	334689-335345	3	49.32	218	22.32	4.77	-	peptide
crov306	336557-337822	2	54.34	421	40.87	5.29	-	peptide
crov307	338258-338070	5	37.57	62	7.08	11.19	-	-
crov308	338535-339425	3	17.06	296	35.82	9.45	-	-
crov309	339469-340665	1	24.73	398	46.62	9.69	-	-
crov310	341558-340662	5	22.52	298	34.43	7.72	1	anchor
crov311	346282-341642	4	20.28	1546	183.08	8.57	-	-
crov312	346361-347446	2	29.74	361	39.44	11.02	-	-
crov313	348474-347491	6	28.76	327	36.92	8.12	-	-
crov314	349442-348504	5	15.34	312	37.51	10.01	-	-
crov315	350014-349493	4	30.46	173	18.40	3.74	2	-
crov316	350079-351512	3	23.71	477	55.44	9.82	-	-
crov317	351616-352539	1	21.86	307	35.97	4.44	1	anchor
crov318	352573-352809	1	27.85	78	9.33	7.34	-	-
crov319	352906-357180	1	21.19	1424	165.57	5.58	-	-
crov320	357246-357806	3	35.65	186	20.00	5.12	3	-
crov321	357880-359841	1	19.01	653	76.63	5.24	-	-
crov322	360517-359834	4	14.62	227	27.56	10.2	-	-
crov323	360889-360527	4	20.94	120	14.23	9.16	-	-
crov324	360983-361339	2	20.17	118	14.21	9.46	-	-
crov325	361386-365852	3	24.87	1488	171.01	9.18	-	-
crov326	365886-367880	3	22.01	664	75.90	5.3	-	-
crov327	367938-368987	3	28.38	349	38.75	4.14	1	anchor
crov328	369519-368950	6	31.40	189	21.88	10.26	-	-
crov329	369640-370590	1	16.30	316	37.77	7.86	-	-
crov330	370595-373840	2	24.03	1081	125.93	9.62	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov331	373889-374419	2	22.22	176	20.53	9.61	-	-
crov332	376288-374489	4	35.11	599	67.33	8.67	-	-
crov333	376370-377659	2	16.74	429	50.49	9.65	-	-
crov334	378096-377668	6	15.85	142	17.04	8.71	-	-
crov335	378474-378157	6	32.39	105	11.72	5.04	1	anchor
crov336	378546-379892	3	24.94	448	51.50	7.87	-	-
crov337	379923-380123	3	20.40	66	7.63	4.04	2	anchor
crov338	380145-380996	3	25.94	283	33.46	9.87	-	-
crov339	381030-381665	3	27.67	211	24.42	7.56	-	-
crov340	381768-382058	3	22.34	96	11.33	4.61	1	anchor
crov341	383146-382100	4	22.35	348	41.95	7.75	-	-
crov342	383416-384936	1	38.33	506	57.24	5.41	-	-
crov343	385038-385607	3	19.65	189	22.56	10.04	-	-
crov344	385633-386262	1	17.14	209	24.67	10.24	-	-
crov345	387026-386271	5	21.43	251	29.85	9.49	-	-
crov346	388971-386983	6	20.31	662	78.35	9.42	-	-
crov347	389255-389731	2	22.85	158	18.33	9.05	-	-
crov348	390293-389958	5	19.64	111	13.04	10.48	-	-
crov349	390367-390999	1	24.49	210	24.25	9.87	-	-
crov350	391050-391277	3	31.58	75	8.47	5.82	-	-
crov351	391684-391292	4	21.88	130	15.69	10.21	-	-
crov352	392208-391780	6	19.81	142	16.97	8.88	-	-
crov353	392279-392950	2	19.20	223	26.84	9.37	-	-
crov354	392984-393418	2	22.76	144	16.09	8.95	-	-
crov355	393809-393420	5	17.18	129	15.29	4.87	-	-
crov356	395131-393881	4	16.87	416	49.70	10.27	-	-
crov357	395212-395829	1	22.82	205	24.04	10.13	-	-
crov358	395964-396956	3	20.14	330	39.45	9.88	-	-
crov359	397001-401692	2	27.71	1563	175.99	4.28	-	-
crov360	401740-402651	1	24.89	303	35.70	4.78	-	-
crov361	402686-403234	2	28.96	182	20.69	7.37	2	-
crov362	403319-403522	2	19.61	67	8.23	10.25	-	-
crov363	404588-403524	5	27.04	354	39.99	4.69	1	peptide
crov364	404674-405417	1	27.28	247	28.65	4.02	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov365	405791-405414	5	21.69	125	14.38	6.65	3	peptide
crov366	406047-405832	6	17.13	71	8.56	8.7	1	anchor
crov367	406117-406485	1	22.49	122	14.12	8	-	-
crov368	406587-411068	3	27.98	1493	169.46	7.34	-	-
crov369	412134-411139	6	18.88	331	39.79	7.53	-	-
crov370	413312-412158	5	25.45	384	43.76	6.94	-	-
crov371	413535-413690	3	25.64	51	6.30	8.48	-	-
crov372	413852-414880	2	34.21	342	38.97	4.28	-	-
crov373	415428-414886	6	17.50	180	21.77	10.12	-	-
crov374	415510-416460	1	27.76	316	36.24	8.14	-	-
crov375	417400-416462	4	20.34	312	37.20	6.07	-	-
crov376	417494-417940	2	20.81	148	17.59	10.4	-	-
crov377	418221-417937	6	18.25	94	10.90	10.83	-	-
crov378	418617-418309	6	16.18	102	12.13	4.88	-	-
crov379	419001-418645	6	18.49	118	13.93	9.92	-	-
crov380	419110-419334	1	25.78	74	8.80	7.69	-	-
crov381	419548-419342	4	26.09	68	8.32	10.53	1	peptide
crov382	421934-419616	5	23.59	772	90.85	6.87	-	-
crov383	422048-424501	2	21.43	817	95.32	8.31	-	-
crov384	426576-424498	6	19.58	692	81.75	8.59	-	-
crov385	426667-426888	1	18.92	73	8.51	6.18	2	anchor
crov386	427680-426892	6	21.80	262	30.46	9.72	-	-
crov387	427770-429185	3	19.63	471	56.31	9.36	-	-
crov388	429234-430265	3	26.55	343	41.00	6.47	-	-
crov389	430319-433498	2	28.84	1059	123.36	7.97	-	-
crov390	433551-436676	3	28.89	1041	119.10	3.78	-	-
crov391	436710-438824	3	31.49	704	81.91	9.2	2	-
crov392	438859-439275	1	14.87	138	16.09	9.81	-	-
crov393	439317-440087	3	19.71	256	30.84	8.5	-	-
crov394	440321-441331	2	21.96	336	39.29	9.86	-	-
crov395	441321-441779	3	23.53	152	17.71	4.7	-	-
crov396	441982-443634	1	24.02	550	64.22	8.35	-	-
crov397	443739-443948	3	21.90	69	8.36	10.21	-	-
crov398	444344-445729	2	24.46	461	53.58	6.91	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov399	445813-446919	1	23.13	368	43.39	4.38	-	-
crov400	446969-448048	2	22.13	359	41.43	6.8	-	-
crov401	448089-448355	3	15.73	88	10.57	9.59	1	anchor
crov402	448399-450849	1	22.97	816	94.40	8.68	-	-
crov403	450896-452842	2	33.59	648	71.21	4.62	-	-
crov404	452920-453339	1	27.86	139	15.54	8.37	2	-
crov405	453783-453346	6	14.84	145	17.66	8.14	-	-
crov406	455558-453816	5	22.15	580	67.12	9.52	-	-
crov407	455634-456242	3	28.41	202	23.20	9.01	-	-
crov408	456835-456443	4	27.99	130	14.93	4.53	-	-
crov409	459029-457050	5	26.01	659	76.99	4.8	-	-
crov410	459094-459423	1	17.88	109	12.71	9.99	-	-
crov411	459888-459433	6	22.15	151	17.90	8.76	1	-
crov412	461065-459908	4	18.83	385	44.76	9.3	-	-
crov413	461156-461590	2	20.46	144	16.92	9.82	-	-
crov414	461627-462061	2	19.31	144	17.39	4.52	-	-
crov415	462249-463568	3	23.56	439	51.09	4.92	-	-
crov416	463809-463573	6	25.74	78	8.94	5.77	-	-
crov417	463880-464365	2	25.51	161	18.12	7.66	-	-
crov418	464419-464673	1	18.43	84	9.85	7.34	-	-
crov419	464693-466153	2	21.29	486	56.65	9.19	-	-
crov420	466965-466228	6	39.70	245	28.40	4.49	-	-
crov421	467829-467125	6	40.28	234	27.02	7.4	-	-
crov422	468726-468019	6	38.42	235	26.53	7.16	-	-
crov423	469357-468884	4	25.53	157	18.80	5.6	-	-
crov424	470831-469401	5	28.79	476	54.27	5.74	1	-
crov425	470894-471292	2	22.81	132	15.13	5.81	4	peptide
crov426	473943-471289	6	22.90	884	103.82	8.57	-	-
crov427	474740-476623	2	24.47	627	72.81	8.56	-	-
crov428	476700-476924	3	22.67	74	8.65	4.52	-	-
crov429	477159-477617	3	25.49	152	16.78	9.35	-	-
crov430	477607-477777	1	32.75	56	5.94	4.31	-	-
crov431	478065-478970	3	17.88	301	36.15	9.53	-	-
crov432	479750-478965	5	20.61	261	30.99	10.25	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov433	480301-479843	4	39.87	152	15.41	6.87	-	-
crov434	480312-480476	3	12.12	54	6.70	10.2	-	-
crov435	480494-483094	2	25.72	866	99.82	4.95	-	-
crov436	483162-483587	3	21.36	141	16.10	9.47	-	peptide
crov437	483634-484332	1	22.17	232	27.15	5.17	-	-
crov438	485046-484531	6	15.89	171	20.65	9.38	-	-
crov439	485129-485701	2	21.64	190	22.36	10.19	-	-
crov440	485743-487842	1	18.67	699	81.98	6.97	-	-
crov441	488511-488344	6	17.86	55	6.43	4.06	-	-
crov442	488608-489945	1	22.94	445	52.42	9.49	-	-
crov443	490242-489946	6	20.88	98	11.63	5.24	-	-
crov444	490767-490276	6	22.56	163	19.29	9.46	1	-
crov445	490837-491628	1	15.40	263	31.11	5.8	-	-
crov446	492521-491625	5	21.74	298	35.56	9.04	-	-
crov447	493967-493365	5	21.39	200	23.40	8.64	-	-
crov448	494266-493991	4	22.83	91	10.63	9.4	-	-
crov449	494453-495085	2	21.80	210	23.82	10.02	3	-
crov450	496098-495127	6	21.60	323	37.18	5.19	-	-
crov451	496775-496143	5	22.43	210	24.94	9.05	1	anchor
crov452	496893-497873	3	27.52	326	38.36	4.95	-	-
crov453	497913-498977	3	25.92	354	40.93	6.36	-	-
crov454	499152-502925	3	25.86	1257	145.30	7.56	-	-
crov455	502975-503436	1	34.63	153	17.19	9.57	-	-
crov456	505273-503438	4	16.61	611	72.26	8.28	-	-
crov457	505372-506442	1	24.46	356	41.73	9.37	-	-
crov458	506495-507547	2	20.70	350	40.98	10.19	-	-
crov459	508332-507535	6	20.18	265	31.13	9.57	-	-
crov460	508425-508844	3	19.76	139	16.40	9.36	-	-
crov461	512859-511855	6	20.60	334	38.36	9.81	-	-
crov462	514858-512903	4	22.65	651	74.84	10.19	-	-
crov463	515128-514919	4	21.90	69	8.15	9.36	-	-
crov464	515514-515359	6	26.28	51	5.96	8.49	-	-
crov465	515909-515556	5	23.45	117	13.09	5.06	-	-
crov466	515964-516566	3	22.39	200	23.42	5.34	1	anchor

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov467	516856-516563	4	20.75	97	11.86	10.24	-	-
crov468	517196-516780	5	22.54	138	16.81	9.57	2	-
crov469	517405-517193	4	19.25	70	8.42	9.2	-	-
crov470	517747-517535	4	23.00	70	8.43	5.56	1	-
crov471	518438-517962	5	18.24	158	19.10	9.52	-	-
crov472	518674-522678	1	22.27	1334	156.32	9.53	-	-
crov473	522720-522893	3	24.14	57	6.84	6.81	-	-
crov474	523063-524187	1	18.84	374	43.94	9.62	-	-
crov475	525014-524205	5	28.89	269	30.50	8.72	-	-
crov476	525462-525046	6	24.46	138	16.19	9.81	-	-
crov477	526518-525523	6	16.47	331	39.76	9.58	-	-
crov478	529688-526560	5	24.03	1042	119.29	9.03	-	-
crov479	529775-531010	2	24.51	411	45.79	10.15	-	-
crov480	531019-532389	1	18.75	456	54.45	8.19	-	-
crov481	532851-532450	6	24.38	133	15.78	9.13	-	-
crov482	532944-533921	3	23.52	325	37.93	5.21	-	-
crov483	534499-533918	4	22.68	193	22.80	6.51	-	-
crov484	534612-535106	3	24.44	164	19.00	10.21	-	-
crov485	535921-535103	4	25.64	272	31.14	5.91	-	-
crov486	538713-535942	6	21.25	923	107.26	9.12	-	-
crov487	539145-538846	6	23.00	99	12.04	10.19	-	-
crov488	539570-539154	5	17.03	138	16.41	10.11	-	-
crov489	539656-540864	1	22.25	402	47.73	7.59	-	-
crov490	541085-540843	5	25.51	80	8.75	4.46	-	-
crov491	541195-541656	1	25.76	153	17.71	4.14	-	-
crov492	542210-541653	5	23.66	185	21.27	9.36	-	-
crov493	542971-542261	4	19.69	236	27.71	9.51	-	-
crov494	543131-545764	2	26.65	877	102.15	7.99	-	-
crov495	545815-547749	1	19.43	644	77.10	9.07	-	-
crov496	547791-548441	3	23.50	216	25.14	4.44	1	anchor
crov497	548497-553314	1	24.57	1605	188.47	8.58	-	-
crov498	553345-557757	1	23.02	1470	173.25	7.6	-	-
crov499	558041-557769	5	22.34	90	10.64	8.4	-	-
crov500	558159-558800	3	24.30	213	24.22	10.07	2	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov501	558822-559412	3	21.83	196	23.20	7.66	-	-
crov502	559466-560824	2	26.64	452	52.36	8.69	-	-
crov503	560881-563493	1	19.79	870	101.57	9.14	-	-
crov504	563510-563845	2	19.64	111	12.96	9.91	-	-
crov505	563890-566757	1	27.68	955	111.60	6.58	-	-
crov506	566947-569160	1	27.82	737	86.49	6.74	-	-
crov507	569236-570519	1	25.23	427	50.48	9.1	-	-
crov508	571535-575170	2	49.28	1211	132.67	8.68	-	-
crov509	576432-575890	6	21.18	180	21.68	8.18	-	-
crov510	577356-577159	6	34.85	65	7.79	10.43	-	-
crov511	577489-578991	1	25.15	500	58.49	9.74	-	-
crov512	579462-580253	2	23.48	263	30.91	9.35	-	-
crov513	580375-580653	1	25.45	92	10.83	10.3	-	-
crov514	580687-581691	1	25.57	334	39.09	8.2	-	-
crov515	581791-584724	1	23.65	977	114.23	8.53	-	-
crov516	584736-585983	3	15.38	415	49.92	9.59	-	-
crov517	586328-586155	5	15.52	57	6.85	10.36	-	-
crov518	586472-586624	2	19.61	50	6.02	4.05	-	-
crov519	586663-587643	1	16.82	326	38.95	9.83	-	-
crov520	587667-588056	3	20.51	129	15.19	10.06	-	-
crov521	588147-588467	3	18.38	106	12.63	9.06	-	-
crov522	588787-588494	4	21.09	97	11.54	9.49	-	-
crov523	588839-589189	2	16.52	116	14.18	9.96	-	-
crov524	589227-589634	3	20.59	135	16.07	9.88	-	-
crov525	589656-591608	3	28.98	650	73.18	6.46	-	-
crov526	594349-592181	4	21.21	722	85.81	9.2	-	-
crov527	595513-597489	1	26.00	658	76.18	5.19	-	-
crov528	597560-598966	2	26.08	468	55.28	9.95	-	-
crov529	599990-600316	2	23.55	108	12.68	4.36	-	-
crov530	600557-601291	2	28.30	244	28.47	5.36	-	-
crov531	601570-601809	1	27.50	79	9.02	4.82	-	-
crov532	602620-603216	1	17.92	198	23.32	8.24	-	-
crov533	603360-604166	3	20.07	268	31.53	7.49	-	-
crov534	604170-604427	3	18.22	85	10.25	9.45	-	-

CDS	Location	Frame	% (G+C)	Length (aa)	MW (kDa)	pI	Trans-membrane helices	Signal peptide or signal anchor
crov535	604887-605975	3	19.65	362	42.92	9.39	-	-
crov536	606022-606312	1	27.49	96	11.71	11.12	-	-
crov537	607161-606328	6	18.35	277	32.30	9.99	-	-
crov538	607228-607398	1	23.98	56	6.49	8.51	-	-
crov539	608018-608890	2	23.37	290	33.91	7.14	-	-
crov540	608976-609170	3	23.59	64	7.58	7.78	-	-
crov541	610230-610958	3	25.38	242	28.06	4.81	-	-
crov542	611279-612271	2	21.25	330	39.25	4.79	-	-
crov543	616905-617327	1	13.00	140	17.45	9.51	4	anchor

Appendix D: List of PCR primers

Table D.1 Oligonucleotide primers used in this study.

Melting temperatures are based on the [4°C per (G,C), 2°C per (A,T)] rule. 'Position' indicates the genome position to which the 5' nucleotide of the primer binds (CroV genome unless stated otherwise); reverse primers are marked by a 'c' after their genome position. Primers that had multiple targets in the genome or did not work for other reasons are not included in this list.

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_1-1	CATCAAAACTACCATTTCCAAC	22	60	435,210c
MF_1-2	TTATTAAAATCGTGCTGCTGG	21	58	444,071
MF_2-1	CAACGGAGTTTATGGTACTAC	21	60	522,417
MF_7-1	TTATAGTACTTATTTAATGGTTC	23	56	25,755
MF_7-2	TTTAACATTTGGGAGTGATTAC	22	58	1,963c
MF_7-3	TATATGTAATCAGAATAGAATATTGC	26	64	23,760
MF_7-4	TAATTTTTGATAAAATAGTGTTGACC	26	64	2,800c
MF_7-5	AAATGCACTTATTATTCTGAACTTCC	26	68	23,379
MF_7-6	CTGTCATCATTCACTATAATAAAAC	25	64	36
MF_7-7	TATATTTTTAATACTTTTAAATTATCAGG	29	66	25,928
MF_7-8	GGATTAAAAATATAAAATTTGAATGG	26	62	2,174c
MF_7-9	GAAGGATTGAAAGAGTTAATTTTAGG	26	68	2,110c
MF_7-10	TTTAAAATACCACCTAATAGTTAAATG	27	66	25,978
MF_7-11	TAAAAATCCATTGCCAGATGAATTAAG	27	70	2,080c
MF_7-12	CTATATAATTTTTAAAATAGATACTC	26	60	25,841
MF_7-14	TATGTCAAAATTAGAGATTACCATCAG	27	70	459c
MF_7-15	AAAATTGATTTTTATAATATTTAAGAG	27	60	500c
MF_7-16	CACTGATGGTAATCTCTAATTTTGAC	26	70	431
MF_7-17	TAATTACTCTTAAATATTATAAAAATC	27	60	468
MF_8-1	TCTATCTACAGGTTGTAATTTC	22	58	213,377
MF_8-2	CTCAACCTCTTGTTTCATAAG	21	58	210,849c
MF_8-3	TAGTTTACAATCATCATATTCG	22	56	211,048c
MF_9-1	CATCATTATCTTCTGTTTCTTC	22	58	209,877
MF_9-2	GATATAGATTTTGAAATTCCTAC	23	58	140,374c
MF_9-4	GAAAGTATGATAGAAACTAATGA	23	58	140,608c
MF_9-5	TGGTAAAGGATGTGAAGGAG	20	58	133,142c
MF_9-6	AATATGAATCATAATTAATGTCTTGC	26	64	213,319
MF_9-7	ATTACAGAAGCAAATACTATTACTCC	26	68	512,511c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_9-8	ATTGGTTGGTTTTTTATTAAAGGAAC	26	66	213,141
MF_9-9	GATGGTAGTTTTACTAAAGTATTGG	25	66	159,647c
MF_9-10	AAAAACAGTTATTATTGAATCTAGC	25	62	158,859
MF_9-11	TCTAAATCTAGTATTCAACAAACAAC	26	66	142,488
MF_9-12	TTAAATAAATCATTTGGTCCATCATC	26	66	144,077
MF_9-13	AAATGAAGTATATCGTATCTTTATGG	26	66	158,560
MF_9-14	TATTGATCATAGAACCACCTTAAAAG	26	68	159,336c
MF_9-15	AATTTTATGTCTTCTGTACTATTATC	26	64	160,441
MF_9-16	TATTCTAATCGTGGTGAAAGACTTGC	26	72	161,243c
MF_12-1	GATTGAAATAGATAAATTACCAG	23	58	289,380c
MF_12-2	ACTTGTTATTGATATAATTCCTG	23	58	305,026
MF_13-1	TTATTATTTCCACATTATTTGATG	24	58	289,190
MF_13-2	GAAGATTTACCTGATAGAACTG	22	60	225,132c
MF_13-3	ATATCACATATTAAAACCCATTC	23	58	288,841
MF_13-4F	GATAATTTCCTGTTGGGGTG	20	58	224,416
MF_13-4R	CACCCCAACAGGAAATTATC	20	58	224,435c
MF_13-5	ATTTTTTTTTGATGGTAAGTTCC	23	58	225,452c
MF_13-6	TAAACTTGAAAATATGGAAAGTC	23	58	223,372
MF_13-7	TAATTTGCTTCTACTACATTATC	23	58	240,871c
MF_13-8	TTTATATGGTGATTTAACAAACC	23	58	240,347
MF_13-9	AATTAAATTCTATTGCCAGGTTTTCC	26	68	214,997c
MF_13-10	GAAGTAAAAAGAATAGAATGGTTACG	26	68	333,748
MF_13-11	ATTTATTTACTAGGGGGTGATATGAC	26	70	214,953c
MF_13-12	AGCTTTTTTACCATATAAGGGTTTTG	26	68	152,842
MF_13-13	AACATTCTGAAAATATTCATAATCAAG	27	66	153,575c
MF_13-14	AATTGAAACCGAAGTTATATTGTAAG	26	66	197,932
MF_13-15	ATCAGTTAATTATCAAAAAATGTTAGG	27	66	198,649c
MF_13-16	TGGATTTATTGGATCACATTTAGTTG	26	68	240,215
MF_13-17	TGGATAATAACCTAAAATAGATTGAG	26	66	214,082c
MF_13-18	TCCTAATAAAATTGCTCCACCAAC	24	66	250,788
MF_13-19	ACTTAACAGGAATGTCTAACGAATC	25	68	251,643c
MF_13-20	ACTTATTTTGTATTAAAAGGTGTTTGG	27	68	276,344
MF_13-21	TATGTGTATCCTCTAGATATAATACG	26	68	277,004c
MF_13-22	ATATTTAATTATTCCTACATATCCTC	26	64	282,384
MF_13-23	TCTCTAACACACCCCTCACCACC	23	72	283,269c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_13-24	ATTATGATGGAAATGAATATTTAGTTC	27	66	315,938
MF_13-25	ATTAGTTTCTTCCCATTCCTTTCC	24	66	316,786c
MF_13-26	GAAACAAATAAAATTCAAGAAGGACG	26	68	319,619
MF_13-27	ATTGAATCTAAATATTGAATAGTTGG	26	64	320,406c
MF_13-28	ACTTAGTGTCTGCTGTTATCATTC	24	66	577,122
MF_13-29	TGGTAATCTTAAAATTATAGGTTCAC	26	66	578,958c
MF_13-30	AGAGATATAGATGTTAGATTAAATG	25	62	202,792
MF_13-31	TTGACTAATATCATAATAATAACTAG	26	62	203,574c
MF_13-32	TCTATTGCTAAATGTAAGGGAGG	23	64	209,681
MF_13-33	TCATAGAGGTAGATTTTTTGATGG	24	64	214,543c
MF_13-34	GATGGATAGGAGTTGTGAATGCG	23	68	335,337
MF_13-35	AAAAACAATACGGTAACCAAGTAG	24	64	575,634c
MF_13-36	TGTTTAAGACTTTGTAAATTAGCCTC	26	68	211,668c
MF_13-37	TAATATTCATCATCAGTATTATTATCTG	28	68	212,253c
MF_13-38	CGAGAAGCAAATGAAGATTTAAGAC	25	68	213,049c
MF_13-39	TAATTAAACCACCTCTTTAAAAAGAC	26	66	212,839
MF_13-40	AATGGCAAACCTAGTAGCAAGTCG	24	70	335,105
MF_13-41F	ACACTGGATGAATCGTCTTACCAAG	25	72	337,942
MF_13-41R	TCTTGCCTCACTGAAATAACTACTG	26	72	337,900c
MF_13-42	CTATCACCTAACAACTCAGTATCTTC	26	72	153,758
MF_13-43	TACTAGGTTTCTTATCAAGTTTCTTAC	27	70	154,669c
MF_13-44	AATGTTTGCTGCTACACTTGTGG	23	66	174,313
MF_13-45	TGCTTCTTGTTCTGTAGGGTCAG	23	68	175,317c
MF_13-46	GAGAGATATGGGTTTTGTTATACAG	25	68	257,746
MF_13-47	GAATTTGAAGATGAACTGGTTGAAG	25	68	258,717c
MF_13-48	ATAATGTTACTAATTATTCAAGATTAAAC	29	66	32,933c
MF_13-49	TCAGGAGTATAGTCTAAATAAGGAG	25	68	182,250
MF_13-50	GCTATATAAAATTGGTCTAATGTTTG	26	64	183,180c
MF_13-52	CTATAATCTTCTTAATAAAAATAGGAC	27	66	31,526c
MF_13-53	CTAAAAAAGATGTAACTGATAATAATC	27	66	32,084c
MF_13-54	ATATAAAAATCATAGTTTAGAAGAATTAC	29	68	30,714c
MF_13-55	GAATCACTTACAGAAATTTATTATGG	26	66	30,743c
MF_13-56	GAACCTAATTCTAGATTTAATAAAGATG	28	70	30,785c
MF_14-1	GTGGTACAGTGTCTAATAGTG	21	60	513,060c
MF_14-2	GGATCTTTATTACTTGCCATC	21	58	517,943

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_14-3	ACGGGAATATGAAGGTTTATATGG	24	66	515,141
MF_14-4	AAACCAGTATATTTTTCAACTACACC	26	68	515,101c
MF_14-5	TAGCTTGTCTTAAATCTCTTTGAG	24	64	512,235
MF_15-1	TATTACATTCTTCAGCATTAAAAACC	26	66	519,142c
MF_15-2	AATTCATCAAGTGACATAAAAAAAGG	26	66	520,702
MF_32-1	GAGGAGTTATGTTATTTGGTG	21	60	536,768c
MF_188-1	ACCAAAAAGTTCAATACCACC	21	58	134,765c
MF_188-2	ATTAAAATAGTTTGATTTATTGCC	24	58	140,169
MF_188-3	CTCCTCACGAACTAATTCTAT	21	58	139,904
MF_188-4	CATCTTCAAGTACCTCACCC	20	60	134,919c
MF_202-3	TGGCTGATGGGCTTAAGTAGTTATTC	26	74	577,378
MF_202-4	ACTCCATATACGCATTCACAACTCC	25	72	335,369c
MF_235-1	CATCTATAATTAATACGCCACC	22	60	305,186c
MF_235-2	AAGGCATCCCATTATTACCAT	21	58	334,211
MF_255-1	TCATCAAACGTATGTGTATAATCACC	26	70	611,817c
MF_255-2	AAATAATGAAATATATACCCACTTTG	26	64	611,764
MF_255-3	ACGAGTAATTAAATGTATTTTATTATG	27	64	611,022c
MF_255-4	AATAAAATAAAGTATTTCTTGGAAAAG	27	64	613,521
MF_255-5	GGTGTATATCAACTACCAGCATGAG	25	72	614,142
MF_255-6	AATTGTGTATCAGATGGAATCATTAC	26	68	614,805
MF_255-7s	TTCCACATTTTATTTAAAAAATCAAC	26	62	614,868
MF_255-8s	CTTCACTTGTTTGTATAACTAAAGAC	26	68	610,257c
MF_255-10s	AGTATAGTCAAAATAATAATTTGATGG	27	66	615,384
MF_255-11s	TAGTTAAAATAATGATTTGATAGTATC	27	64	615,313
MF_265-1	AGCACACATAGTGAAAAAATG	21	56	595,978
MF_265-2	GTTACCAAAATTTAAAGTATATAA	24	56	595,377
MF_288-1	CTTTAACAGCACTTACAATTAG	22	58	458,578c
MF_288-2	CGTCTGTTTGTTGGGTAATTG	21	60	467,631
MF_290-1	CCGTAGGTCACTCGTTCG	18	58	509,053
MF_290-2	TGGAGATATGTTTAGTGATATG	22	58	495,265c
MF_290-3	CTACGGAACTACAGTCCGCC	20	54	509,058c
MF_290-4	TTTATACTATCCTGATGTCGGAAAC	25	68	508,411
MF_290-5	TGTGCTCAAATTATATGTTTCCG	23	62	508,451c
MF_290-6	TTGGTTCAATATCTACCCAATGT	23	62	507,678
MF_290-7	TCACTACCCGTAAAATACAATAATGC	26	70	507,384c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_290-8	GATGAGTTAAAAGGATATACTGCTCC	26	72	506,756
MF_290-9	AATGTTTCATATCAAGTTCTAATGGA	26	66	506,203c
MF_290-10	ATCAAAATGCTGTATTGCATCC	22	60	505,421
MF_290-11	AGATGTTAATGGTAATACTCCCTACC	26	72	505,115c
MF_290-12	AAAGGATTAGAACCTGTAATGGGA	24	66	504,564
MF_290-13	TCTTCCAACTGAAATGTCTTTAACA	25	66	504,119c
MF_290-14	CGTAGCAAACCATCAGCAAA	20	58	502,806
MF_290-15	TCCTAATCCAATAGGACGATGTCT	24	68	501,920c
MF_290-16	AAACCAAACTGTATTTATTGTAAATGG	27	68	501,579
MF_290-17	TCTGTAAAGTGTCCTGAAGAAGGT	24	68	500,830c
MF_290-18	GGTTTAATTTTAGGTAAAGGCACG	24	66	500,421
MF_290-19	AATGCGACCAGCTAATAATGC	21	60	499,439c
MF_290-20	TGTGTCTGTGGATGGTACACAA	22	64	498,461
MF_290-21	ATGGATCCCTTCCATCATCA	20	58	497,222c
MF_290-22	TGTTTTAGTTATTGGTTCAGTTTCA	25	62	495,703
MF_290-23	ATATCTTTAATATGGTCTGCATGTCC	26	70	508,163
MF_290-24	ACTTTTCATTAGAAGCACGTAAAGCC	26	72	496,219c
MF_291-1	GAAATATGGAAAATACTACTATG	23	58	538,907c
MF_291-2	GTTAATATAGTAATTGAAGATGG	23	58	475,539c
MF_291-3	TAGGTAAATTAGACAAATCATG	22	56	476,510c
MF_291-4	ATATTAACATTTTTATTACTTGTTCC	26	62	522,266c
MF_291-5	TATAATAAAGAAACAGATATTGTTCC	26	64	424,950c
MF_291-6	AAGTTAATAGAGTCTGTTAAATCACG	26	68	480,219
MF_291-7	AATTTTAGTGATATTATTTATACAGC	26	62	557,713
MF_291-8	ATGTTTGTGATATTATTATACAAGAG	26	64	458,980c
MF_291-9	TAACTGTTCAATAGGTGAATTTTGG	25	66	491,292c
MF_291-10	TCATTGAATTGATATCTTTACGTGTG	26	68	490,209
MF_291-11	AAGTAGAAGATGAACTAGGACAAC	24	66	490,001c
MF_291-12	ATGAAAATTATCAAGAACAAGAAGC	25	64	489,323
MF_291-13	ATTGGGTAATTTTACAGTTATTGGG	25	66	488,658c
MF_291-14	ATTAGATATAAATAAATGGATACAAGG	27	66	487,761
MF_291-15	AAATCCTTCTAATGGTTCTAAAATG	25	64	484,711c
MF_291-16	ATGTGAAAGCTGTATAAGTGAAGG	24	66	483,521
MF_291-17	TATTTTAACTTGTATGGGTATTAACG	26	64	466,504c
MF_291-18	AATGTAAAGGTTCTTAGTGAATATGG	26	66	465,755

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_291-19	AAGCAGATTTAGATGATGAGGATG	24	66	457,171c
MF_291-20	TACCAGTATAGTTAAAGGAGATGG	24	66	456,008
MF_291-21	TTACTTTTGATAGACATATGACAATG	26	66	455,461c
MF_291-22	TTTTTTGGTTCATTATTACCTATAAG	26	64	453,928
MF_291-23	TTTAAGAAGAAAACAACCATCAGATG	26	66	441,746
MF_291-24	AATTTTATGGGGTCATATTCAAAGG	25	64	441,010
MF_291-25	AATTGGTGGTAACGATGAAGATAAG	25	68	436,912c
MF_291-26	TTTGGTTACTACAAGTGATAATCGG	25	68	436,298
MF_291-27	CTTAGTTGAATCACAAAAAGCACC	24	66	435,413c
MF_294-2b	CTACATTTTTATCTAACAATTTATC	25	60	214,450c
MF_299-1	TGAAGAAACAGTTCCCATCC	20	58	416,553
MF_299-2	GTTGGAACAAATTGATTAGCC	21	58	399,591c
MF_301-1	ACTGGGTGATGATTGTAATAC	21	58	602,571c
MF_301-3	GTGACCATATTTTAAGTGTATC	22	58	607,697
MF_301-4	TAAAACTTCTAAAAAATCATGTGACC	26	66	607,677
MF_301-5	CTAGAACAAGATGAGGTAATATATCC	26	70	602,861c
MF_301-6	AATGGAATCTAAAAATCCTGATATGC	26	68	606,237
MF_302-1F	CTGGTAACACATTTACAATATG	22	58	599,873c
MF_302-1R	CATATTGTAAATGTGTTACCAG	22	58	599,852
MF_304-1	TAAACGATATTGTTTGAGCATC	22	58	523,083c
MF_304-2	TCTATACCATAACAGTTAGAAC	22	58	536,267
MF_304-3	ATATGATGTGTTTCTAATTGTTG	23	58	536,108
MF_304-4	CTATGATTCGGAAATCCCATC	21	60	538,466
MF_305-1F	GGCATATTATTAGATATTTGTGA	23	58	600,561c
MF_305-1R	TCACAAATATCTAATAATATGCC	23	58	600,539
MF_306-2	CCATATTATGTTCCTCCTTTAC	22	60	133,038
MF_309-1	AGTGAAGCTGGATTGCATCAAAG	23	66	571,552c
MF_309-2	AAAACGTATGGTATTAACTCAAGGTG	26	66	569,511
MF_309-3	GAACTTCCTGTCCTCACCACTGTAG	25	76	573,818
MF_309-4	GTTCACTTTCTTCTCGCCTGGACG	24	74	572,275c
MF_311-1	TTTCCTGATTCTTTCATATTGC	22	58	71,263c
MF_311-2	AATCTAAGAATTGATTAAACCCC	23	60	110,491
MF_311-3	GAAGATTGGTTAATGGTTATAG	22	58	110,001
MF_311-4	GTTTTAATAGGCAAATTATCTAAT	24	58	110,415
MF_311-5	GATTAAACCCCTCACCAAAAG	21	60	110,503

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_311-6	TAACCTTTAATTTTTTAGTTTGTCC	25	62	71,922c
MF_311-7	CTTGTAAAATCCACCTTTAATCAACC	26	70	132,528
MF_311-8	TACTATTTTATTAAATTCTAAACCTC	26	62	72,311c
MF_311-09	AATTATAAGGACTTTTAATTTTTAAC	26	60	110,407c
MF_311-10	TTAATAACAAGGTCTTTCTTTGCTAG	26	68	44,451
MF_311-11	ATGTAAAAACAATTTTCACCAGCATG	26	68	45,322c
MF_311-12	ATATTTCATGCCATACCACTTTAG	24	64	59,172
MF_311-13	AGAAACATCTCAACAAATTAGATATG	26	66	59,698c
MF_311-14	ACTGAAGATAATCCTTCTAATAATG	25	64	78,965
MF_311-15	AATCTATGGGATTAACACAGTTTG	24	64	79,702c
MF_311-16	TAGAAGAAATAAAAGAACCTATTCC	25	64	107,847c
MF_311-17	ATATGAAAGTGTGTTGCTCTGTG	23	64	115,576
MF_311-18	ACCTGATAAACAACAAAAATATACTG	26	66	116,440c
MF_311-19	TATCTAAACGACGAACATATTTTG	24	62	120,470
MF_311-20	ACATTTGTTAATAATAAATCAACTGG	26	64	121,333c
MF_311-21	ATCTTGTGATAATTTAGGTATAGG	24	62	131,092
MF_311-22	TTTATAATTTTTTGATATAATACACGG	27	64	131,815c
MF_311-23	TTTGATAGACCTCAAGAAATAGG	23	62	131,862
MF_311-24	GTGAACTGGTAGATATAAGTGG	22	62	133,200c
MF_311-25	TGTTGTATCATCTTTATTATCACTC	25	64	42,506
MF_311-26	AACAAAATTGGAAATGAATTATATTC	26	62	43,326c
MF_311-27	TTACCATTAAGGAGTCAGAACCTG	24	68	90,636
MF_311-28	AAATAAGGATGGCTATTATATTCCAG	26	68	91,134c
MF_311-29	AAATTAAAAATACCGAAATTAGGTC	25	62	122,324
MF_311-30	AGTATTTCAACCTATTAATTGGATC	25	64	124,661c
MF_311-0kb_fwd	GGATGTATATCAATTAGAGAAG	22	58	113,049c
MF_311-3kb	GAAGATTGGTTAATGGTTATAG	22	58	110,001
MF_311-4kb	CTCTTAACGTAGGAGATACG	20	58	109,027
MF_311-5kb	TATTTTAACATACCCATCTTTAC	23	58	108,032
MF_311-6kb	CATATACTAACAATTTACGCCC	22	60	107,047
MF_311-7kb	TTTACCATATAATCCATTCAAAC	23	58	106,052
MF_312-1	CATGTGTTTTTAGATTATAATCAGAAAC	28	70	29,257c
MF_312-2	AAATTAATCATTTTTTATTTAGTTTCTG	28	62	29,209
MF_312-3	CTACAATTTCTTGATTAAACCTATATC	27	68	30,095
MF_319-1	CTAATTTATCAAGATGAGTTGC	22	58	213,733c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_320-1	TTATTTTCCCTAATGCAGGTG	21	58	416,778c
MF_320-2	TGTAAGTTATCAAGATTTAGAAC	23	58	423,418
MF_320-3	AGAATCAGAAACAAATACATGG	22	58	416,996c
MF_323-1	CAATTATTTAGCTGAAATGGGT	22	58	133,307
MF_323-2	GTATTGAAACTGACCTAATTTG	22	58	134,242
MF_324-1	ATTCAGAACAAGCTAAATTGCTCCC	25	70	337,994
MF_324-2	TCCAGTGTGAAACACGTTTCCCC	23	70	337,949c
MF_326+-1	ACACCAACTTATAATTGTACAG	22	58	474,755
MF_326+-2	GCCAGTGATAAATGTATGAAAG	22	60	468,946c
MF_328-1	GCAATATACACACAATTACATG	22	58	580,733
MF_331-1	CTTTATCTAAAAAATCCACAATG	23	58	424,418c
MF_331-2	GGAGATCAAATTGACGGAAG	20	58	434,433
MF_332-1	GAGTTTCTTATATTATCCATCC	22	58	558,178c
MF_332-2	CAGGTTTTGATGACTCTCTTG	21	60	569,795
MF_332-3	CAAATGATGACCAAAAAACAAC	22	58	558,354c
MF_332-4	GAATTAACAGGATTGGTCAAAC	22	60	569,066
MF_332-5	AAGTTTATTATTTATACTCAATTTCC	26	62	568,810
MF_332-6	ATTCATTGTTGTCGTCATAGGTAGCC	26	72	558,909c
MF_332-7	TACCCTACACTCATAGATATAGTGC	25	70	558,769c
MF_332-8	AAACAGAAGTGGATGAAGATTGTCC	25	70	557,447
MF_333-1	CTAACAAGTATAGTAAATGGAC	22	58	599,386
MF_333-2	CAATCTAAACTGCTTGGTAAAT	22	58	597,392
MF_334-1	CCAGTTCTATTACATCTTTATC	22	58	455,347c
MF_334-3	TAATTTCTTGTATACTGGAAGG	22	58	457,217
MF_335-1	ATCCAACAGAAACACTAATCG	21	58	399,318
MF_335-2	ATGTGTTATTTCTTGAGTAACC	22	58	338,558c
MF_335-2b	ATAGAACCTATAAATGTTAAAGG	23	58	338,731c
MF_335-2c	TCCTAAGATACGAAAATTATGG	22	58	337,230c
MF_335-3	GGTTATGACATACCAGTAATG	21	60	342,601c
MF_335-4	ATAGTTTTATTACTAATGTATCTCC	25	62	422,961
MF_335-5	AATCAGTGTCAATCTCTTTTTGAACC	26	70	338,901c
MF_335-6	AAAGATTATGCACCTGTTAAAAATGC	26	68	391,612c
MF_335-7	TTTCCAATAGATAATAGTGAAAAAGC	26	66	390,682
MF_335-8	AGAGCATTCTCTATAATTTGGATGG	25	68	399,642c
MF_335-9	ATTTAATAATGAGAATAAATACATGGG	27	66	399,090

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_335-10	GGTAATGCTCCCAAACTATCTCTG	24	70	398,616c
MF_335-11	TTAAACTATAGAACTACTTTTGGAGG	26	68	398,135
MF_335-12	GTGTGGTTTGTTTAATACCAGGAG	24	68	379,149c
MF_335-13	TGAAGGTATTGAATCTATTTATCAGG	26	68	378,617
MF_335-14	TAATCCTGTTTTAGAACATTGGAAAG	26	68	347,716c
MF_335-15	GTTCTTTATCCCTGAACAATCTGG	24	68	346,615
MF_335-16	CCAGATTGTTCAGGGATAAAGAAC	24	68	346,638c
MF_335-17	TTTTTGCCTGGTTTTTAGATAACTC	25	66	345,955
MF_335-18	TATTTTTTCCATAACTTCCATTACG	25	68	338,811c
MF_335-19	AAGACAGCACCTATGTCATGTGG	23	68	574,868
MF_335-20	TTTGTTATTTCATGTGTTATTTCTTGAG	28	70	338,569c
MF_335-21	TAGTTCCGTAGGTCACTCGTTCG	23	66	509,048
MF_335-22	TATATCTGATACTATTCAAAGTAGATG	27	64	512,404c
MF_335-23	ACGCAGTTCCAAGCATTTGTTCAAG	25	72	574,477
MF_335-24	GTATTGACTTATGTAGAACATTTATAG	27	68	575,948c
MF_335-25	TGATGTCGGAAACATATAATTTGAG	25	66	508,423
MF_335-26	TTAGTAAAAATACTGATATAGTTGTGG	27	68	513,039c
MF_335-27	ATCAGAGAAACACTTCCAACAAAGG	25	70	574,644c
MF_335-28	GCAACACGGAAGATGGTAGAGTG	23	70	573,728
MF_335-30	ATGGATTATTTAGAACCTCATTTAAG	26	66	507,518
MF_335-31	AATATTATTTATTTTTAGTATTATTTAC	28	60	508,893
MF_335-33	GATGATTTAGGGATGAGTGTGTCG	24	70	508,809
MF_335-34	TAGGGATGAGTGTGTCGCCACC	21	68	508,816
MF_345-1	GTATTCATACTTTAAGATTTGG	22	56	594,604
MF_345-2	ATTTTACATAATCCAATGGCTG	22	58	581,512c
MF_345-3	ATAGGTTCTCAAACGTGAATG	21	58	594,248
MF_345-4	CATTTGTAAGTTGTATAAGTTGAC	24	60	581,816c
MF_345-5	TACAAGTTAGATTTATGTTGTTTGCC	26	68	598,817
MF_345-6	TTTTAAATGTGTAGGTACTTTATGCC	26	68	580,614c
MF_345-9	ATTGAAAGTTTAACACAAATTATTGG	26	64	606,765c
MF_345-10	TCTATTAATCCTCCATTTATTTCTGC	26	68	605,872c
MF_345-11	ATAACAAATTTAATTTTCAAGGAACG	26	64	605,089
MF_345-12	AACCTATAATTTTAAGATTACCACC	25	64	578,936
MF_345-13	ATCAACCACTAGATTTCTTGCCCG	24	70	597,296
MF_345-14	ATAATATCATCATATGCACTTTTTGG	26	66	598,911c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_345-15	ATCACCAATAATTTGTGTTAAACTTTC	27	68	606,736
MF_345-16	CGTAACTTTATTGAATATAATCGTATGC	28	72	606,271
MF_345-17	TGTTTGTATAAAATGAGATGGAACACTC	28	74	606,427
MF_345-18	TATTAACTTGTAGTAAAACTTTCGATC	27	68	607,286
MF_345-19	ATAACCTAAATTACTAGTTGATATGAC	27	68	607,865
MF_345-20a	TAAATAAAATAAATGATAGT	20	44	609,022
MF_345-20b	TGAAGATTTAAATAAAATAAATGATAGT	28	64	609,014
MF_345-21s	ATACCTTAACTTTTGACTCCTATTTTG	27	70	609,046
MF_348-1	GTTTTTGTTTTCTTGTTGATGG	22	58	444,242c
MF_348-2	TCGTTTATTAAAGTTGCCACC	21	58	455,103
MF_350-1	GACTCATCCATAATTTGATATAAATCG	27	70	616,911c
MF_350-2	CGATTTATATCAAATTATGGATGAGTC	27	70	616,885
MF_350-3s	CATGTTGATGGTTTATTTTAATATATC	27	66	616,631c
MF_350-4	ATTTAATATTTAGGATCATGTTGATGG	27	68	616,647
MF_354-1	GTACATCATTTCCTATATCTTC	22	58	480,911c
MF_354-2	GATATTCATTAACATTTAGACAC	23	58	50,870c
MF_354-3	GTAAACATAGTGATGATGAAAG	22	58	51,083c
MF_354-4	TGGCGGAGAGGATGATTTAC	20	60	57,572c
MF_354-5	CCTTGTTGTGGGTTCCATCT	20	60	56,550
MF_354-6	ATTTCACATCCAATTGCTCCAGAACC	26	74	481,545c
MF_354-7	TTAGTTTTTGGTCTTGGTAAACAACC	26	70	31,563c
MF_354-8	ACACAAAATAGTCTAATTGATAATGG	26	66	70,866
MF_354-9	ATATCACTTATTAGAGAAATTCCTGG	26	68	494,780
MF_360-1	GGTGTTTGATAATACCATTGAT	22	58	50,663
MF_360-2	GATGTGAATTAGAATATGTTCC	22	58	31,259c
MF_360-3	CATATAATCTTGATGTGCTTGA	22	58	50,197
MF_299-335_fwd	ATGCTATTCCTGTTTTAAGAAC	22	58	399,737c
MF_299-335_rev	ATACATGGGATTCAAACAGTG	21	58	399,108
MF_9-337_fwd2	GCCACAATGTTGTACGGAAAT	21	60	206,745c
MF_9-337_rev	TTACGATATTTTTGATCTCCTC	22	58	205,509
MF_11-326_fwd	GGCTTATCAAGTATCATCAAC	21	58	473,440c
MF_11-326_rev	TTTCTAATAAACTAACTCCTCC	22	58	472,734
MF_13-294_fwd	GTAAGTTCCAATAGGTATCAG	21	58	225,438c
MF_13-294_rev	TGATTTGTGTTGGTTGATAATG	22	58	224,361
MF_265-345_fwd	TAATTCTTTTAAGCTCCCAGG	21	58	595,425c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_265-345_rev	ATTCCTGTTTGATAGGTTCTC	21	58	594,237
MF_328-333_fwd	TCATTTGAGTTTTGGAGATAG	21	56	581,001
MF_328-333_rev	CGGGCAAGAAATCTAGTGG	19	58	597,319c
MF01	AATTTGTATGAACGTCGAGAG	21	58	104,031c
MF02	CCTGATTAATTGCTATTCGTC	21	58	103,605
MF01_10kb	GACCCTAATCCAACAAGATG	20	58	94,003
MF01_15kb	TACTGCCAATTATCATAGAACC	21	60	89,007
MF01_20kb	AGTTGTTGCTTCAACCAGAG	20	58	84,114
MF_polBint_fwd	TCAGTAATTGGAGATACTCC	20	58	550,894
MF_polBint_rev	AATTATAGAACCTACACCTGC	21	58	551,925c
MF_MCP_fwd2	CTCGTGGCTTATGGTGCTCAAGAC	24	74	383,437
MF_MCP_rev2	TACACGAAGAACATTGTAGTT	21	56	384,900c
MF-pSMART193F2	CATCATCAAACATATATTCACTC	23	60	170,923
McG-PL01-04F1	TGAGGATGAACTCTTTTCCC	20	58	172,398c
McG-PL01-04R1	GTACTTGTGGAAGTGATTTTAG	22	60	172,779c
McG-PL01-06F1	TCAAGTGACTACCATGAACTG	21	60	57,447c
McG-PL01-06R1	GTACTCCTGAAGCTCATGTG	20	60	56,980
McG-PL01-07F1	ATATGACATATCTACAATTTTTAG	24	58	570,348c
McG-PL01-07R1	TTCTACCAATAATTTGTTTTTCG	23	58	569,068c
McG-PL01-27F1	CTGGCTTTTTAATTGATCCAAC	22	60	225,132
McG-PL01-30F1	GTATATAATTTTGTTGTTCAGGG	23	60	435,087c
McG-PL01-30R1	TATAACACTGATAACCAATTTTC	23	58	434,046
McG-PL01-51R1	GGAACAATTCTACATCTACATG	22	60	434,392
McG-PL01-86F1	TCCACCTCCTGCACAAAATG	20	60	224,278c
McG-PL01-96F1	TATAAGAGCATTCATAAATTGAG	23	58	223,873
McG-PL01-07F2	TGTGAAGGTAATGGAACAAAC	21	58	569,738c
McG-PL01-27F2	AACAGTATTGTCTTCATTGGTA	22	58	225,082
291++_331-F	GAATCACAAAAAGCACCAGG	20	58	435,406c
291++_331-R	ATTATCATCTGGTGGGATTGA	21	58	434,315
McG-PL01-86F2	ATGTACAATTTTTTCCAGGAAC	22	58	223,937c
13-repeat-3kb-F	CAGTTCATTGTGTAAATGATTTGGAA	26	68	225,605c
13-repeat-3kb-R	CCACTAATAGAAATTGCTCAAAGCAG	26	72	222,648
13-repeat-5kb-F	ACTCTTCCCAAGGTTCATATGGATTA	26	72	226,573c
13-repeat-5kb-R	TTTTGATGTTTGGATAACAACTGACC	26	70	221,635
13-5kb-F	TGTTGATTCCTGAAACGACAGAACTA	26	72	249,410c

Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
13-5kb-R	GATGAATATTCCGCTAGTGATAATTGG	27	74	244,389
13-3kb-F	ATCCGTCCATATTACTTCATTTTCCA	26	70	248,361c
13-3kb-R	AACCCAAAGTGTTGTATTCCCATTTA	26	70	245,285
MF_FLX94-1	GTAAAATTACGTTTAAGATTGAATATAG	28	68	27,589c
MF_FLX94-2	CTATATTCAATCTTAAACGTAATTTTAC	28	68	27,562
MF_tRNA-2	TATTTTGTTAATGGTTTAAATTCTAAATC	29	68	509,995
MF_tRNA-3	TATCTTTCTAAATTAGGAGATTTAG	25	62	510,041c
MF_tRNA-4	CAAGTATAATGTTGTTGCTTTACCC	25	68	510,105c
MF_tRNA-5	GGGTAAAGCAACAACATTATACTTG	25	68	510,081
MF_tRNA-6	GAAGTAATCTTATGTCACTTGTAGG	25	68	510,147c
MF_tRNA-7	CCTACAAGTGACATAAGATTACTTC	25	68	510,123
CroV-qPCR-1	CCAATCAACAGCTACATTAAC	21	58	12,436c
CroV-qPCR-2	GTTGGTAAAAGTGGTATTATTC	22	58	12,317
CroV-qPCR-3	CTGTCGGACAAGGCACAC	18	58	171,781
CroV-qPCR-4	CTAGTAAGATGGGATGTTC	19	54	171,952c
CroV-qPCR-5	GATACTACAGGGTTAGCAG	19	56	375,180
CroV-qPCR-6	GTATATGGACGTTATGATGG	20	56	375,325c
CroV-qPCR-7	GAGCTATGGCAAGACCTATG	20	60	590,810
CroV-qPCR-8	TGTCTTTTAAAGTTCTCTGGG	21	58	590,937c
MF_351-1	GTTATTGTTGAATACAGAAGTTCATG	26	68	GSFLX_82 / GS20_351: 521c
MF_351-2	CACTGTTCCATTTAATTGTTAAAGGC	26	70	GSFLX_82 / GS20_351: 580
MF_351-3	AACTAGATATTCAAGAATCTCATAAAG	27	68	GSFLX_82 / GS20_351: 682
MF_FLX79-1	CGCCCAATCGGTAGCCCAATC	21	68	GSFLX_79: 626
MF_FLX79-2	GATTGGGCTACCGATTGGGCG	21	68	GSFLX_79: 646c
MF_Spezl-01	TGCTTCTTATGCTTGAAATGTTCC	24	66	Mavirus: 427c
MF_Spezl-02	TATCGTGGTATATTTTGTGATTCTC	25	66	Mavirus: 337
MF_Spezl-03	AAGCGGTGGTTAATACGGACAC	22	66	Mavirus: 13,174
MF_Spezl-04	TAGTTTCATTAAGCCAAATATACTG	25	64	Mavirus: 13,880c
MF_Spezl-05	TGAAAGCGACAGTGATGATGAAG	23	66	Mavirus: 1,719
MF_Spezl-06	TAAAATATCAAGGATTTGAAGAGG	24	62	Mavirus: 2,177c
MF_Spezl-07	TTCGCTTAAAACTTGAGACCGTCC	24	70	Mavirus: 1,416c
MF_Spezl-08	CACAGGATATTTCAGAACCAACGC	24	70	Mavirus: 17,810
MF_Spezl-09	TGTATTTAACAATTTAGGAGATTCAG	26	66	Mavirus: 18,242

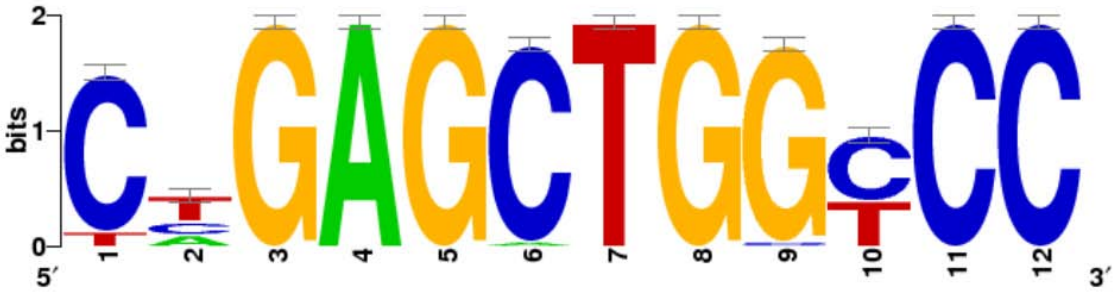
Primer name	Sequence 5'→3'	Length	Tm (°C)	Position
MF_Spezl-10	GTTTTATTTGGGTTTCCTTCTATAAG	26	68	Mavirus: 347c
Spezl-qPCR-1	ATGGTGCGACTTTTACTTATG	19	58	Mavirus: 4,042
Spezl-qPCR-2	AAGGGGTTATATTTTCCATTAG	22	58	Mavirus: 4,201c
Spezl-qPCR-3	GATTGGTATTGCGGATGATG	20	58	Mavirus: 15,586
Spezl-qPCR-4	TCTAATGAAATTAACGGGGTG	21	58	Mavirus: 15,750c
MF_307-1	TGTGCTGATAATTGATCGTTC	21	58	phage307: 43,648
MF_307-2	CTCTATATCAATCTGATCTGC	21	58	phage307: 564c
MF_307-2b	CGTCTCTTTACGCAAGCATC	20	60	phage307: 635c
MF03	CGAAACGACCTCTGATCGC	19	60	phage307: 38,406c
MF04	GCACATAAAGCAGATCGCAG	20	60	phage307: 37,852
MF-pSMART193R2	GGCAATGGCAGGAAATGAAG	20	60	phage307: 10,201c
MF-pSMART194R2	CCGTTAAACCATTTATCAGATC	22	60	phage307: 11,613c

Appendix E: The genome of phage307

During 454 GS 20 pyrosequencing of the CroV genome, a single 43 kbp contig (contig 307) stood out from the rest of the dataset, mainly with regard to its GC content (56% compared to 23% for the rest of the contigs), but also by its gene content. The majority of GS 20 contigs exhibited CDSs with high similarity to protistal and Mimivirus genes, but the CDSs on contig 307 were most similar to bacteriophage and bacterial genes. PCR with primer sets oriented distally at the ends of contig 307 resulted in a single product, and sequencing confirmed the circular nature of this 43,651 bp long bacteriophage genome. The same genome was found again in the 454 GS FLX dataset, although it was split into three contigs. The host for phage307 is unknown, but it can be assumed that it is part of the mixed bacterial assembly present in *C. roenbergensis* cultures. PCR primers specific for the phage genome yielded a product when tested on genomic *C. roenbergensis* DNA isolated from CroV-uninfected cultures. PCR with CroV-specific primers on the same template was negative. Phage307 is thus unlikely to be associated with CroV. However, it is not clear why this bacteriophage co-purified with CroV in sucrose and Optiprep™ density gradients.

The missing sequence information in contig 307, which was later filled in by PCR and Sanger sequencing, was 287 bp long and contained 34 copies of a 12 bp direct repeat with the consensus sequence shown in Fig. E.1.

Figure E.1 Consensus sequence of the terminal direct repeat in phage307.



To define start and end of the putative phage genome, the direct repeats were arbitrarily split in half, with the first 17 copies of the repeat assigned to the start and the remaining 17 copies to the end of the genome. In a preliminary genome analysis, I annotated 59 ORFs as putative protein-coding sequences. The genome organization of phage307 is shown in Fig. E.2. BLASTP analysis revealed that several CDSs had close homologues in bacteria (mostly prophages) and bacteriophages (Table E.1). Putative functions could be assigned to some of them, including a Holliday junction resolvase, a terminase, a phage-type integrase, a DnaJ heat

shock protein, an endonuclease VIII, an antirepressor, a capsid protein, a capsid protease, a portal protein, and a tail fibre protein. Two phage307 structural proteins (CDSs 31 and 32) were identified during the LC-MS/MS study of the CroV virion proteome (see section 3.3.9). The best BLASTP hits of the predicted structural proteins were to podoviruses infecting *Burkholderia*, *Escherichia*, and *Salmonella*. Furthermore, the genome size of 43,651 bp fits the size range of podoviruses (40-42 kbp). It is therefore likely that phage307 belongs to the family *Podoviridae*. In order to assess the ecological impact of this bacteriophage, further experiments will have to be done, first and foremost to identify the bacterial host.

Figure E.2 Genome diagram of phage307.

CDSs on forward and reverse strands are shown in blue and the G+C content is represented by the inner black circle (relative to the genomic average of 55.63%).

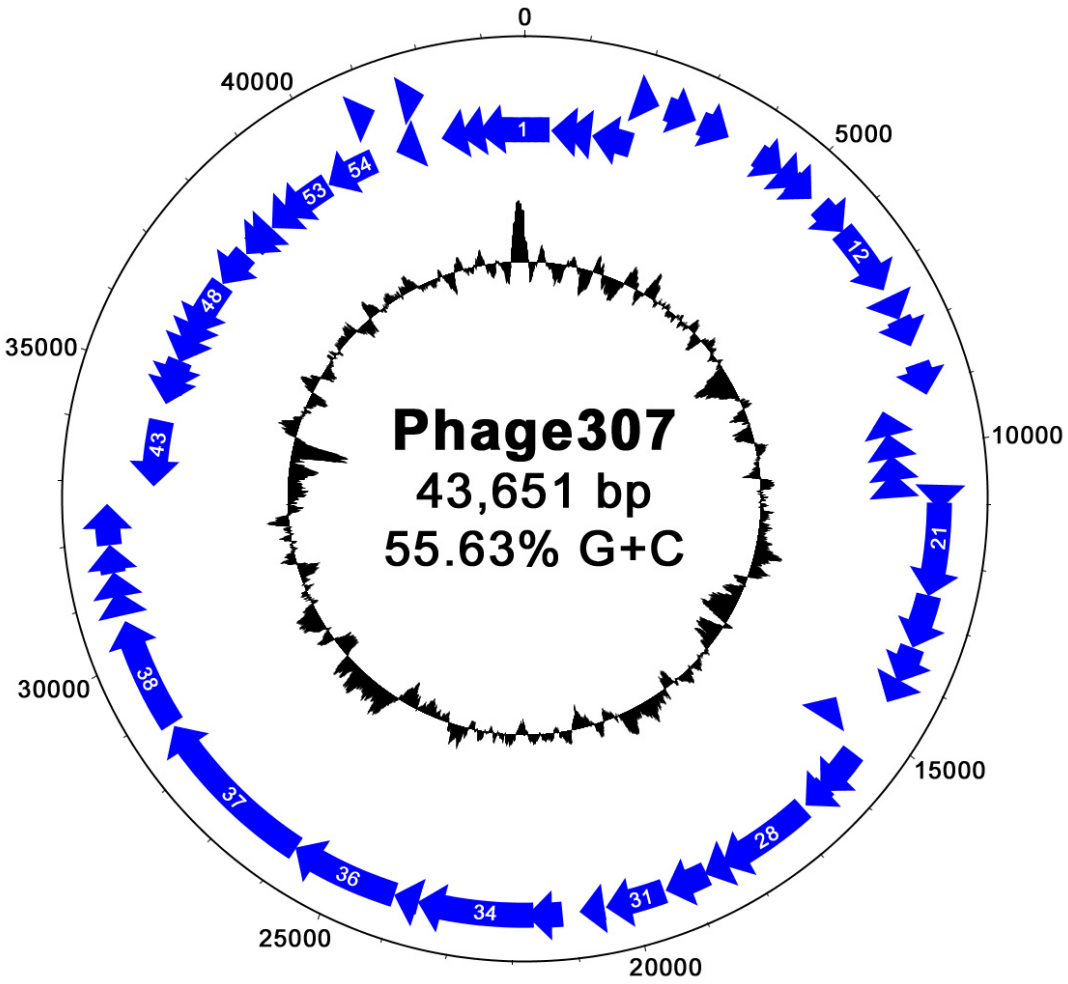


Table E.1 Putative CDSs of phage307 and their annotation.

CDS	Strand	Genome location	Length (aa)	Predicted function/comments	Top BLASTP hit	E-value	Percent identity	Alignment length
1	-	441-42785	435		*Xanthobacter autotrophicus*	$7x10^{-78}$	41	463
2	-	789-502	95					
3	-	1070-786	94					
4	-	2044-1277	255	phage repressor, peptidase, HTH domain	*Magnetospirillum magnetotacticum*	$1x10^{-29}$	39	193
5	+	2075-2296	73	signal peptide	*Ahrensia sp*. R2A130	$2x10^{-4}$	38	68
6	+	2445-2954	169	AP2 domain, HNH endonuclease, transcription factor	*Enterobacter cloacae*	$3x10^{-30}$	48	159
7	+	3056-3565	169					
8	+	4142-4633	163	putative DNA segregation ATPase FtsK/SpoIIIE	*Bordetella petrii*	$8x10^{-16}$	41	107
9	+	4636-5004	122	endodeoxyribonuclease RusA, Holliday junction resolvase	*Burkholderia sp.* Ch1-1	$1x10^{-25}$	52	136
10	+	4859-5293	144					
11	+	5492-6055	187					
12	+	6052-7245	397	P-loop NTPase	*Rhodobacter sphaeroides*	$2x10^{-49}$	38	350
13	+	7511-7798	95					
14	+	7795-8283	162					
15	+	8608-9132	174					
16	-	9488-9228	86					
17	-	10051-9659	130					
18	-	10399-10064	111					

CDS	Strand	Genome location	Length (aa)	Predicted function/comments	Top BLASTP hit	E-value	Percent identity	Alignment length
19	-	10806-10396	136					
20	+	10760-11074	104					
21	+	10977-12527	516	terminase	*Acinetobacter baumannii*	$1x10^{-116}$	45	516
22	+	12530-13426	298					
23	+	13438-14058	206					
24	+	14184-14432	82					
25	-	15144-14809	111					
26	+	15469-16290	273	antirepressor	*Ahrensia sp*. R2A130	$9x10^{-33}$	50	130
27	+	16191-16637	148		*Nitrosomonas sp.* AL212	$3x10^{-10}$	36	105
28	+	16709-18349	546	portal protein	*Klebsiella pneumoniae*	$3x10^{-57}$	31	498
29	+	18349-18678	109					
30	+	18659-19390	243	capsid protease	*Treponema denticola*	$3x10^{-5}$	29	171
31	+	19431-20426	331	major capsid protein, found in proteome	Enterobacteria phage epsilon15	$6x10^{-107}$	55	334
32	+	20465-20872	135	found in proteome	*Paracoccus denitrificans*	$3x10^{-22}$	44	144
33	+	21170-21754	194	gp11	*Burkholderia multivorans*	$2x10^{-15}$	32	171
34	+	21654-23591	645		*Desulfovibrio salexigens*	$4x10-^{66}$	31	558
35	+	23591-24004	137		*Opitutaceae bacterium*	0.002	37	98
36	+	24008-25834	608					
37	+	25834-28722	962		*Phaeobacter gallaeciensis*	$5x10^{-15}$	23	633
38	+	28786-30660	624	tail fiber protein, lectin	*Chthoniobacter flavus*	$1x10^{-12}$	35	148
39	+	30860-31180	106		*Roseovarius nubinhibens*	$9x10^{-12}$	43	92
40	+	31217-31501	94					

CDS	Strand	Genome location	Length (aa)	Predicted function/comments	Top BLASTP hit	E-value	Percent identity	Alignment length
41	+	31494-31964	156		*Pseudoalteromonas atlantica*	$6x10^{-31}$	49	150
42	+	31968-32648	226					
43	-	34221-32962	419	phage integrase	Candidatus *Puniceispirillum marinum*	$1x10^{-58}$	33	412
44	-	34775-34527	82		Enterobacteria phage Eco1230-10	0.094	32	73
45	-	35404-34772	210	DnaJ	*Mesorhizobium sp.*	$3x10^{-47}$	50	192
46	-	35624-35355	89					
47	-	35960-35649	103					
48	-	37042-35939	367	gp37/gp68	*Burkholderia sp*. CCGE1002	$7x10^{-65}$	43	382
49	-	37758-37039	239		*Burkholderia sp.* Ch1-1	$6x10^{-32}$	39	223
50	-	37967-37755	70					
51	-	38359-37964	131	Transcriptional Regulator, XRE family	*Magnetospirillum gryphiswaldense*	$2x10^{-20}$	48	115
52	-	38746-38441	101					
53	-	39783-38746	345	P-loop NTPase	*Methylobacterium nodulans*	$1x10^{-73}$	52	279
54	-	40760-39780	326	endonuclease VIII, gp42/gp60	*Gluconacetobacter diazotrophicus*	$5x10^{-44}$	35	304
55	+	40832-41101	89					
56	-	41486-41193	97					
57	+	41624-41962	112					
58	-	42415-42083	110					
59	-	42770-42426	114		*Xanthobacter autotrophicus*	$5x10^{-9}$	39	119

Appendix F: CroV in the media

Following the publication of the CroV genome in *Proceedings of the National Academy of Sciences USA* [32], news agencies around the world reported about the largest marine virus in online and print media. The following represents a selection of online articles from different countries.

"Giant virus may be ocean's largest"
http://news.sciencemag.org/sciencenow/2010/10/scienceshot-giant-virus-may-be.html?rss=1

"Giant marine virus found"
http://www.the-scientist.com/news/display/57776/

"More complicated than you think - A new, giant virus is confounding old certainties"
www.economist.com/node/17358573

"Giant virus with tiny victims carries a monster genome"
http://www.newscientist.com/article/dn19625-giant-virus-with-tiny-victims-carries-a-monster-genome.html

"Newfound virus is giant with lifelike properties"
http://www.msnbc.msn.com/id/39837713/ns/technology_and_science-science/

"Giant marine virus discovered in ocean"
http://www.cbsnews.com/8301-501465_162-20020669-501465.html

"World’s largest, most complex marine virus is major player in ocean ecosystems: UBC research"
http://www.science.ubc.ca/news/478

"Forscher staunen über schwimmendes Riesenvirus" (Germany)
http://www.spiegel.de/wissenschaft/natur/0,1518,725254,00.html

"Forscher entdecken Riesenvirus im Meer" (Germany)
http://www.focus.de/wissen/wissenschaft/natur/groessen-rekord-forscher-entdecken-riesenvirus-im-meer_aid_565622.html

"Riesenvirus im Meer identifiziert" (Switzerland)
http://www.20min.ch/wissen/news/story/18555008

"Největší mořský virus" (Czech Repbulic)
http://www.rozhlas.cz/leonardo/zpravy/_zprava/nejvetsi-morsky-virus--802661

"Il virus gigante che viene dal mare" (Italy)
http://www.galileonet.it/articles/4cc7dfcf72b7ab3ea7000005

"Le plus gros virus marin jamais découvert" (France)
http://www.futura-sciences.com/fr/news/t/biologie-3/d/le-plus-gros-virus-marin-jamais-decouvert_25790/

"Descubren el virus marino más grande y complejo del mundo" (Spain)
http://www.plataformasinc.es/Noticias/Descubren-el-virus-marino-mas-grande-y-complejo-del-mundo

"Vírus gigante" (Brazil)
http://www.diariodasaude.com.br/news.php?article=virus-gigante&id=5902

"World's largest marine virus is 'major player in ocean ecosystems'" (India)
http://news.oneindia.in/2010/10/26/worldslargest-marine-virus-is-major-player-in-oceanecosy.html

"Tutkijat löysivät uuden jättiläisviruksen" (Finland)
http://www.hs.fi/ulkomaat/artikkeli/Tutkijat+l%C3%B6ysiv%C3%A4t+uuden+j%C3%A4ttil%C3%A4isviruksen/1135261173253

"Биологи изучили ДНК морского вируса-гиганта" (Russia)
http://www.lenta.ru/news/2010/10/26/giant/

"种来自海底捕食性浮游生物的巨病毒" (China)
http://www.biotech.org.cn/news/news/show.php?id=82187

"Phát hiện chủng virus lớn, phức tạp nhất thế giới" (Vietnam)
http://www.baoangiang.com.vn/newsdetails.aspx?newsid=24966

〝海は巨大ウイルスでいっぱいだ〞(Japan)
http://d.hatena.ne.jp/semi_colon/20101103/p1

The CroV PNAS article has been selected and evaluated by the Faculty of 1000:
http://f1000.com/5910956