6-2012

Paramecium bursaria Chlorella Virus 1 Proteome Reveals Novel Architectural and Regulatory Features of a Giant Virus

David Dunigan
University of Nebraska-Lincoln, ddunigan2@unl.edu

Ronald Cerny
University of Nebraska-Lincoln, rcerny1@unl.edu

Andrew T. Bauman
Ocean Biologics, Seattle, WA

Jared C. Roach
Institute of Systems Biology, Seattle, WA

Leslie C. Lane
University of Nebraska-Lincoln, llane1@unl.edu

See next page for additional authors

JVI Accepts, published online ahead of print on 13 June 2012
J. Virol. doi:10.1128/JVI.00907-12

Title: *Paramecium bursaria* Chlorella Virus 1 Proteome Reveals Novel Architectural and Regulatory Features of a Giant Virus

David D. Dunigan[#] (1, 2), Ronald L. Cerny (3), Andrew T. Bauman (4), Jared C. Roach (5), Leslie C. Lane (1), Irina V. Agarkova (1, 2), Kurt Wulser (3), Giane M. Yanai-Balser (1), James R. Gurnon (1), Jason C.Vitek (1), Bernard J. Kronschnabel (1), Adrien Jeannard (6), Guillaume Blanc (6), Chris Upton (7), Garry A. Duncan (8), O. William McClung (8), Fangrui Ma (2), James L. Van Etten[#] (1, 2)

[1]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE 68583-0900, USA;

[2]Nebraska Center for Virology, University of Nebraska-Lincoln, Lincoln, NE 68583-0900, USA;

[3]Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304, USA;

[4]Ocean Biologics, Seattle, WA 98133, USA

[5]Institute of Systems Biology, Seattle, WA 98103, USA;

[6]Structural and Genomic Information Laboratory, UMR7256 CNRS, Aix-Marseille University, Marseille, FR-13385, France

[7]Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada;

[8]Department of Biology, Nebraska Wesleyan University, Lincoln, NE 68504-2794, USA

Running title: PBCV-1 virion proteome

Key words: Chlorella, Phycodnaviridae, large dsDNA virus, protein mass-spectrometry

Abstract: 174 words

Text: 7,017 words

#Corresponding authors: David Dunigan, email: ddunigan2@unl.edu and James L. Van Etten, email: jvanetten1@unl.edu

ABSTRACT

The 331 kilobase pairs chlorovirus PBCV-1 genome was re-sequenced and annotated to correct errors in the original 15 year old sequence; forty codons was considered the minimum protein size of an open reading frame. PBCV-1 encodes 416 predicted protein encoding sequences and 11 tRNAs. A proteome analysis was also conducted on highly purified PBCV-1 virions using two mass-spectrometry based protocols. The mass spectrometry-derived data were compared to PBCV-1 and its host *Chlorella variabilis* NC64A predicted proteomes. Combined, these analyses revealed 148 unique virus-encoded proteins associated with the virion (about 35% of the coding capacity of the virus) and one host protein. Some of these proteins appear to be structural/architectural, whereas others have enzymatic, chromatin modification and signal transduction functions. Most (106) of these proteins have no known function or homologs in the existing gene databases except as orthologs with other chloroviruses, phycodnaviruses and nuclear-cytoplasmic large DNA viruses. The genes encoding these proteins are dispersed throughout the virus genome and most are transcribed late or early-late in the infection cycle, which is consistent with virion morphogenesis.

INTRODUCTION

Complex cellular and viral processes are modular and accomplished by the concerted action of functional modules. One of the important functional modules of a virus is the virion particle, which ranges in complexity from a single type protein and small nucleic acid (e.g., tomato bushy stunt virus) to having dozens of types of proteins and lipids, along with a large nucleic acid genome (e. g., poxviruses). Regardless, whether "simple" or complex in composition, all virions carry the legacy of their progenitors through encapsidation, release and stabilization. Virions facilitate propagation of progeny through a series of tightly regulated biochemical steps, called the immediate-early phase of infection, which includes attachment, penetration, uncoating of the viral genome, intracellular trafficking of the viral genome to its replication center, and augmentation of cellular functions to "accept" the exotic nucleic acid/replicon. The architectural elements of virions tend to be prominent, but studies on the supergroup nucleocytoplasmic large DNA viruses (NCLDV) (7, 37, 43) indicate that in addition to structural components, these virions perform multiple enzymatic and regulatory functions that are partitioned among several proteins. The purpose of this study was to determine the virion proteome of *Paramecium bursaria* chlorella virus 1 (PBCV-1), a member of the NCLDV (11, 54).

PBCV-1 is the type member of the genus *Chlorovirus* (family *Phycodnaviridae*) that infects certain chlorella-like green algae from fresh water sources; these viruses are found throughout the world (54, 56). The chlorovirus host algae are normally symbionts of aquatic protists, and in this state are resistant to virus infection. Nevertheless, virus titers from natural sources have been measured as high as 10^5 plaque forming units (pfu)

per ml; however, titers fluctuate with the season (58, 61). Very little is known about the role chloroviruses play in freshwater ecology (41), but susceptible hosts lyse within 6-16 hours in the laboratory and burst sizes typically exceed 10^2 pfu per cell (54, 56). Thus, chloroviruses have the potential to alter microbial communities both quantitatively and qualitatively, as well as act as a driving force for microbial evolution (11). Fortunately, some of the host algae can be grown in the laboratory independent of their co-symbiotic protists.

The 331 kilobase pair (kbp) PBCV-1 dsDNA genome was sequenced and annotated about 15 years ago (26) and reported to have 689 open reading frames (ORF) of at least 65 codons. Of these 689 ORFs, 377 were predicted to encode proteins (CDS); PBCV-1 also encoded 11 tRNAs (reviewed in 21, 55, 57). The size of PBCV-1 extends beyond its coding capacity; the virion is a T = 169d quasi-icosahedral particle with a diameter of 190 nm across the 5-fold axis (63, 64) and has an estimated molecular mass of greater than 1 x 10^9 Da (53). The virion is ~64% protein consisting of at least 40 polypeptides, as seen on one dimensional SDS-PAGE (42). The particle contains 5-10% lipid, which is associated with a bi-layered membrane underneath an outer glycoprotein shell (5, 42, 64).

The capsid structure consists of the major capsid protein (MCP, Vp54), which is glycosylated at 6 sites (31) and is myristylated at least at one site (36). Vp54 complexes with itself, and perhaps other proteins, to form homotrimeric capsomers that are responsible for the planar features of the capsid. Initially it was assumed that, except for the 12 vertices, Vp54 was the only protein contributing to the external capsid and 5040 copies of Vp54 were predicted per virion (64). However, recent studies indicate that the

PBCV-1 virion is more complex than previously thought. i) PBCV-1 contains a unique vertex with a 560-Å-long spike structure, which protrudes 340 Å from the surface of the virus. The part of the spike structure that is outside of the capsid has an external diameter of 35 Å at the tip, expanding to 70 Å at the base. The spike structure widens to 160 Å inside the capsid and forms a closed cavity inside a large pocket between the capsid and membrane enclosing the virus DNA (5, 66). The related chlorovirus CVK1 has a virion-associated protein Vp130 (homolog of PBCV-1 A140/145R) that binds to algal cell walls and is located at a unique vertex (34, 35), suggesting this protein is associated with the spike structure. ii) Regularly spaced appendages occurring on the surface of the virion are present at approximately 1 per trisymmetron (66). These appendages probably assist in attaching the virion to its host cell (56). iii) The volume of the capsomers at the common vertices and those surrounding the spike structure at the unique vertex differ significantly, suggesting they consist of different proteins (5, 66). iv) At least one vertex region may have a retractable appendage, such that when probed with a scanning atomic force stylet the structure retracts, but then resets much like a plunger with a spring (23). It is not known if this plunger is at the unique spike structure vertex or one of the other 11 vertices. v) Six minor capsid proteins of varying stochiometries support the particle architecture and appear to interact with the internal membrane in both the tri- and pentasymmetron structures, as observed with an 8.5 Å resolution map of the virion (66). Of these, a “long protein” (~32 kDa) with similarity to the PRD1 bacteriophage long glue proteins forms an hexagonal network over the internal surface of the trisymmetrons, and a “membrane protein” dimer (~28 kDa) is located at the edge of the trisymmetrons and is connected to the internal membrane (1, 8). vi) PBCV-1 DNA binding proteins

were evaluated with proteomic methods from isolated viral DNA of virions (59). Six proteins were identified that have high isoelectric points that are well suited for binding and neutralization of DNA. Thus, PBCV-1 structure has both symmetric and asymmetric elements, adding to the complexity of the virus morphology.

vii) In addition to these structural features, PBCV-1 contains several functions that initiate infection. PBCV-1 attaches specifically to its host *Chlorella variabilis* NC64A. Thus, we predict one or more surface proteins of the virus mediate attachment; it is probably the spike structure (66). Immediately upon PBCV-1 attachment, the cell wall is degraded at the site of attachment. viii) Virions contain cell wall degrading activity (28, 62). ix) Within the first minutes of infection the cell membrane depolarizes (12, 32), leaving the cell with significantly altered secondary transporter functions (2). This activity is hypothesized to be partially due to a PBCV-1-encoded K^+ channel, Kcv (A250R) (27); however, no direct evidence supports the presence of Kcv in the virion. x) In the first five min of infection host DNA begins to degrade and this is likely due to the two virus-encoded DNA restriction endonucleases [R.CviAI (A579L), R.CviAII (A252R)] packaged in PBCV-1 virions (3). Host chromatin degradation begins before viral transcripts appear. PBCV-1 DNA is resistant to the restriction enzymes because it is methylated. xi) The next major intracellular event is the synthesis of early viral transcripts, observed 5-10 min p.i. (67; Blanc et al., unpublished data), which apparently occurs by pirating the cellular transcriptional machinery, because the virus does not encode a recognizable RNA polymerase gene and no polymerase activity was detected in virion-derived extracts (Jon Rohozinski and James Van Etten, unpublished results).

The purpose of the current study is to evaluate the total viral complement of

proteins associated with the PBCV-1 virion using proteomic technologies and to re-examine the structural/architectural features of this virus, as well as the initial events of infection in the context of the protein complement. This evaluation led to the re-sequencing of the PBCV-1 genome after preliminary proteomic analyses suggested there were errors in the PBCV-1 genome sequence (26). This report presents the newly revised PBCV-1 genome and annotations, and proteomic analyzes of the infectious particles.

MATERIALS AND METHODS

Virus, cells, and culture conditions. Procedures for growing virus PBCV-1 in the alga *C. variabilis* have been described (3, 52, 53).

Virus purification scheme. The virus was purified essentially as described (52) with the following modifications. Prior to sucrose density gradient separation, the virus-cell lysate (2 liters) was clarified by incubating with 1% (v/v) NP-40 detergent at room temperature for 1 - 2 h with constant agitation followed by centrifugation in a Beckman Type19 rotor at 53,000 ×g, 50 min, 4°C. The pellet fraction was solubilized in virus storage buffer (VSB) (50 mM Tris-HCl, pH 7.8) and layered onto a 10 - 40% (w/v) linear sucrose density gradient made up in VSB, centrifuged in a Beckman SW28 rotor for 20 min at 72,000 ×g at 4°C. The virus band was identified by light scattering, removed from the gradient and concentrated by centrifugation. Resuspended virus was incubated with 50 µg/ml proteinase K in VSB for 4 h at 25°C to disassociate and degrade contaminating proteins (this treatment has no effect on virus infectivity). The proteinase K treated virus was layered onto a 20 - 40% linear iodixanol (OptiPrep™, Axis-Shield, Oslo, Norway) gradient in VSB and centrifuged at 72,000 ×g in a Beckman SW28 rotor for 4 h at 25°C for isopynic separation. The gradient produced a single major light-scattering band at

~32% iodixanol corresponding to a density of 1.171 g/ml. The virus band was removed by side-puncture of the centrifugation tube, diluted approximately 10 fold with VSB, then concentrated by centrifugation in a Beckman Ti50.2 rotor at 80,000 ×g for 3 h at 4°C. The pellet fraction was re-suspended in VSB, then filter sterilized with a 0.45 µm cutoff membrane, and stored at 4°C. The virus was quantified by UV/visible scanning spectroscopy using an extinction coefficient of $A_{260/0.1\%} = 10.7$ (52) and plaque assayed to determine the number of infectious particles. These preparations typically yielded several milliliters of stock virus at 1 - 10 x 10^{11} pfu/ml. The infectious to total particle ratio is normally 0.25 - 0.5 for such preparations (53).

These preparations were used both for re-sequencing the PBCV-1 genome and the determination of the proteome; the proteome was determined by two independent methods using mass spectrometry of trypsin digested proteins.

Re-sequencing and annotation of the PBCV-1 genome. Preliminary proteomic analyzes using the existing PBCV-1 gene annotations (NCBI Refseq: NC_000852) revealed possible errors in the genome sequence, which prompted us to re-sequence the PBCV-1 genome. PBCV-1 DNA was purified from virions treated with DNase I, sequenced using Roche 454 Life Sciences GS FLX Titanium chemistry, and assembled as described in the Supplemental Information section (SI). PBCV-1 contigs were identified and annotated as described in the SI.

Proteomics method 1. SDS-PAGE/Trypsin/HPLC/Ion Spray/MS-MS.

Particle disruption and protein extraction. The PBCV-1 virion proteome was evaluated with two independent methodologies, see Figure 1. In the first method virion proteins were solubilized essentially as described (25) with reduction of the proteins by

adjusting 50 µg of virions in 50 µl. An equal volume of cracking buffer [50 mM Tris pH 8.5, 5 mM of the reducing agent dithiothreitol (freshly reduced with tributylphosphine; in some experiments beta-mercaptoethanol was substituted for dithiothreitol), 1% SDS, 0.1% crystal violet and 1% Ficoll 400] was added. The sample was heated to 100^{o}C for 3 min. The reduced proteins were subsequently alkylated by adjusting the solution to 12.5 mM iodoacetamide with a 0.25 M stock, then heating to 100^{o}C for 1 min. These samples were immediately subjected to SDS-PAGE. Alternatively, the proteins were alkylated without previous reduction by the same procedure.

Alternatively, phenolic extractions were used to isolate virion proteins. Reduced and alkylated proteins were adjusted to 40% sucrose to increase the density of the solution. These preparations were then extracted with an equal volume of water-saturated phenol or water-saturated phenol with toluene added to increase the hydrophobicity of the phenol. The protein-containing phenolic phase was removed, and protein was precipitated with 10 volumes of methanol then dissolved and heated in cracking buffer.

One-dimensional SDS-PAGE. Proteins were separated on thirty-two cm linear gradient (4-20%) polyacrylamide gels with 0.1% SDS and 375 mM Tris, pH 8.7 tank buffer of 25 mM Tris/190 mM glycine. The samples were electrophoresed at room temperature till the crystal violet tracking dye reached the bottom of the gel.

The gel was fixed and stained with Sypro-Ruby according to the manufacturer's recommendation (Life Technologies Corporation). The stained gel was imaged using a blue box transluminator. Once imaged, the gel was cut into 32 one cm size pieces being careful to clean the scalpel between samples. These gel pieces were then processed for

trypsin-digestion and mass spectrometry analyses.

MS-based microsequencing. Excised gel pieces were digested for peptide sequencing using a slightly modified version of a method described by (40). Briefly, the samples were washed with 100 mM ammonium bicarbonate, reduced with 10 mM DTT, alkylated with 55 mM iodoacetamide, washed twice with 100 mM ammonium bicarbonate, and digested *in situ* with 10 ng/µl trypsin. Peptides were extracted with two 60 µl aliquots of 1:1 acetonitrile:water containing 1% formic acid. The extracts were reduced in volume to approximately 25 µl using a vacuum centrifugation.

Ten µl of the extract solution was injected onto a trapping column (300 µm x 1 mm) in line with a 75 µm x 15 cm C18 reversed phase LC column (LC- Packings). Peptides were eluted from the column using a water + 0.1% formic acid (A)/95% acetonitrile:5% water + 0.1% formic acid (B) gradient with a flow rate of 270 µl/min. The gradient was developed with the following time profile: 0 min 5% B, 5 min 5% B, 35 min 35% B, 40 min 45% B, 42 min 60% B, 45 min 90% B, 48 min 90% B, 50 min 5% B.

The eluting peptides were analyzed using a Q-TOF Ultima tandem mass spectrometer (Micromass/Waters, Milford, MA) with electrospray ionization. Analyses were performed using data-dependent acquisition (DDA) with the following parameters: 1 sec survey scan (380-1900 Da) followed by up to three 2.4 sec MS/MS acquisitions (60-1900 Da). The instrument was operated at a mass resolution of 8,000. The instrument was calibrated using fragment ion masses of doubly protonated Glu-fibrinopeptide.

Mass ion analyses. The MS/MS data were processed using Masslynx software (Micromass) to produce peak lists for database searching. MASCOT (Matrix Science,

Boston, MA) was used as the search engine. Data were searched against the NCBI non-redundant database. The following search parameters were used: mass accuracy 0.1 Da, enzyme specificity trypsin, fixed modification CAM, variable modification oxidized methionine. Protein identifications were based on random probability scores with a minimum value of 25. Although this number varied from experiment to experiment, typically it was 25 or less for $p < 0.05$ confidence.

Relative abundances. Approximate, relative quantitation of the proteins was determined using the exponentially modified protein abundance index (emPAI) (18). This method uses the number of observed peptides compared to the number of observable peptides giving a ratio that is directly proportional to relative abundance of the protein in the mixture when adjusted exponentially ($emPAI = 10^{PAI}-1$; where PAI = number of observed peptides per protein/number of observable peptides per protein).

Proteomics method 2. PPS/Trypsin/HPLC/MS-MS

Protein extraction and trypsin digest. One hundred µg of PBCV-1 was mixed 1:1 with 100 mM ammonium bicarbonate buffer pH 8.3 containing 0.2% PPS (Protein Discovery Labs, San Diego, CA) [final concentration 50 mM ammonium bicarbonate, 0.1% PPS], boiled for 5 min, cooled to room temperature, reduced and alkylated with 5 mM dithiothreitol and 15 mM iodoacetamide, then digested with sequencing grade trypsin at a 1:50 trypsin:protein ratio, for 4 h at 37°C, with shaking. The digested samples were acidified with HCl (200 mM), incubated at 37°C, and centrifuged at 4°C, to remove PPS prior to LC-MS application.

LC Methods. Buffer solutions were made with LC-MS grade water, acetonitrile, and formic acid and consisted of 5% acetonitrile/0.1% formic acid in water (Buffer A)

and 100% acetonitrile/0.1% formic acid (Buffer B). Two or 4 µg total protein from each sample was loaded onto a reverse phase (RP) trap (5 µm, 200 Å, Magic; Michrom Bioresources, Auburn, CA) with 100% buffer A and washed for 10 min prior to separation on a microcapillary column. The microcapillary column was constructed by slurry packing 18 cm of C18 material (2.7 µm, 100 Å, HALO, Michrom Bioresources) into a 75 µm ID fused silica capillary, which was previously pulled to a tip diameter of 5 µm using a Sutter Instruments laser puller (Sutter Manufacturing, Novato, CA). Separations were performed on an Eksigent 1D+ nano-LC (Eksigent, Dublin, CA) LCQ-Deca XP Plus: 0-30% B over 240 min, 30-70% B over 10 min at 300µl/min; LTQ-Velos: 0-30% B over 80 min, 35-70% B over 10 minutes at 300µl/min.

Mass Spectrometry Methods. Data-dependent tandem mass spectrometry (MS/MS) analysis was performed using an LTQ-Velos or LCQ Deca XP Plus mass spectrometer (ThermoFisher, San Jose, CA). Full MS spectra were acquired in centroid mode, with a mass range of 400–2000 Da. To prevent repetitive analysis, dynamic exclusion was enabled with a LTQ-Velos: repeat count of 1, a repeat duration of 30 sec, an exclusion list size of 500, and an exclusion-duration of 90 sec. Tandem mass spectra were collected using a normalized collision energy of 35% and an isolation window of 3 Da.

For the LTQ one full scan was followed by 6 MS-MS scans of the 6 most intense precursor ions not on the dynamic exclusion list. LCQ-Deca XP Plus: repeat count of 1, a repeat duration of 30 sec, an exclusion list size of 100, and an exclusion-duration of 20 sec. Tandem mass spectra were collected using a normalized collision energy of 35% and an isolation window of 4 Da.

For the LCQ one full scan was followed by 3 MS-MS scans of the 3 most intense precursor ions not on the dynamic exclusion list.

Mass ion analyses. Processing and searching of MS/MS spectra and analyzing peptide and protein identification data were performed using SPIRE (Systematic Protein Investigative Research Environment, www.proteinspire.org) system with default parameters. Searches were conducted using the X!Tandem search engine (9) within a 2.5-Da mass error, a variable modification for methionine oxidation (16@M), and a fixed modification for iodoacetamide (57@C) along with the default search parameters. The sequence file for the searches of the modules contained PBCV-1 appended to a decoy database of *Ostreococcus tauri*. In addition, a randomly reshuffled version of each database was appended for error estimation. The search results were processed with the LIPS (logistic identification of peptide sequences) model (16) to generate peptide spectra scores. Peptide identification probabilities and FDRs were calculated based on the reshuffled matches using an isotonic regression model (17). A 90% certainty was used as the basis for spectra identifications. A recently introduced approach was used to estimate the protein identification FDR from individual peptide identification probabilities (17).

RESULTS AND DISCUSSION

Re-sequenced and re-annotated PBCV-1 genome. The original sequence and annotation of PBCV-1 was completed over 15 years ago using primitive procedures when compared to current technology. During the past 15 years we have corrected the sequence of individual genes as mistakes were detected. Those mistakes and preliminary results from the current proteomic analyses that indicated sequencing errors, prompted us

294 to re-sequence PBCV-1. The revised PBCV-1 genome contains 330,805 nucleotide pairs
295 compared to 330,743 nucleotide pairs from the earlier sequencing effort. The two
296 genome versions differed by 458 indel positions (mostly single nucleotide indels) and
297 188 substitutions. This genome sequence and annotation are deposited at the National
298 Center for Biotechnology Information (NCBI) as reference sequence NC_000852.5; the
299 genome annotation is listed in SI Table S1. The re-sequenced genome submitted to
300 NCBI includes the 2,222 base pair terminal inverted repeat ends, but not the incompletely
301 base-paired covalently closed hairpin 35-nucleotide loops at each end of the genome.
302 Thus, the genome is a linear double-stranded DNA of 330,805 base pairs with two 35-
303 nucleotide partially paired terminal loops. Sequencing reads were obtained through the
304 hairpin loops (data not shown). When compared to the published results from Zhang et al.
305 (1994), the terminal repeats and hairpin loops are identical. Nucleotide 1 refers to the first
306 paired nucleotide following the hairpin loop.

307 One significant change in the new annotation is that ORFs of 40 codons or more
308 were classified as potential CDSs; the previous annotation used 65 codons as the
309 minimum size. This resulted in 802 ORFs, of which 416 ORFs were classified as
310 "major" CDSs (designated with an upper case "A") based on the following supporting
311 evidence: these ORFs did not have larger overlapping ORFs and/or were expressed
312 transcriptionally (65) and/or the protein was identified in the proteomic analyses. The
313 major ORFs cover 92.8% of the genome sequence and have an average protein product
314 size of 249 amino acids. In addition, 11 tRNA genes were identified as reported
315 previously. The remaining 386 ORFs were labeled "minor" ORFs (designated with a
316 lower case "a") and most of them are probably not CDSs. They encode putative proteins

317 with an average size of 86 amino acids. The gene annotations, along with functional
318 assignments, are listed in SI Table S1.
319
320 To avoid confusion in the literature, we kept the same gene numbering system as
321 used previously, i.e., a gene labeled as *a250r* is still labeled *a250r*. When two adjacent
322 ORFs were found to be a single ORF, e.g., A189R and A192R, we named it A189/192R.
323 Finally, where smaller ORFs were identified that were not considered previously, we
324 labeled them with a lower case letter, e.g., A254aR. These new gene annotations were
325 used for the proteomic analyses of the virion proteins.
326 **PBCV-1 virion proteome.** Highly purified virions were used for the proteome
327 analyses, including a "protease treatment" step where the particles were incubated with
328 proteinase K to degrade proteins non-specifically associated with the particle surface.
329 Proteinase K treatment does not affect PBCV-1 infectivity (3). Using a combination of
330 sample treatment, separation and mass spectrometry methods, 148 virus-encoded proteins
331 were detected in the PBCV-1 virion (Fig. 2B). For abundant proteins, any method was
332 sufficient to detect mass ions allowing identification with high confidence. However,
333 some of the low abundance and small proteins were only identified by one of the two
334 methods, primarily due to differential separation where the protein of interest was
335 separated from an abundant, and consequently masking, protein. The dynamic range of
336 these analyses was ~10^4 with the MCP present at approximately 10^3 copies per virion
337 relative to a hypothetical protein present at one copy per virion. Thus, the sample
338 treatment and separation method selected were important elements in the proteome
339 determination. The proteins were identified by two independent methods, 62% of the
340 proteins were detected by both methods. Twenty six percent were uniquely identified

with the SDS-PAGE method (Method 1) and 11% were uniquely identified with the PPS solubilization method (Method 2). It is important to note some proteins are not readily detected using mass spectrometric methods, e. g., small proteins associated with membranes (39). Thus, the proteome reported here may increase with additional data in the future. However, the results presented are the compilation of many experiments using varying conditions for protein extraction and isolation giving us high confidence in the compiled list of proteins including several proteins with predicted transmembrane domains, as well as many small proteins; i. e., less than 10 kDa (Table 1).

Method 1. SDS-PAGE/Trypsin/HPLC/Ion Spray/MS-MS. Method 1 identified 137 virus-encoded proteins in the virion. Virion proteins were either: i) extracted directly into gel sample buffer, ii) first extracted into a phenolic phase to remove nucleic acids, or iii) extracted into a hypo-polarized phenolic phase supplemented with toluene to further extract highly polar proteins such as glycosylated proteins. The extracted proteins were either alkylated with iodoacetamide and then reduced, or left alkylated. While these methods helped extract certain proteins, others were excluded and no additional proteins were detected beyond the standard method of extracting into the gel sample buffer.

Protein separation using one-dimensional gel electrophoresis resolved ~30 distinct SYPRO-Ruby stained bands. The dynamic range of observed polypeptides is large. For example, the MCP migrates at approximately 54 kDa and is the most abundant protein in the virion, migrating near the mid-point of the gel (Fig. 2A, gel position 13). The MCP has a nominal mass of 48 kDa and is post-translationally modified with sugars at 6 positions (31) and with at least one myristyl group (36), as well as having the amino

terminal methionine removed (13). This very abundant protein contrasts to proteins detected in regions of the gel where little or no staining was observed, e. g., gel positions #1, 8, 9, 31, 32 in Fig. 2A. Although very little staining was observed in these regions, several proteins were detected by the mass spectrometry analyses. Indeed, proteins were detected in all regions of the gel.

Qualitative changes in protein mobility were observed with different sample treatments (SI Fig. S1). Samples that were alkylated with iodoacetamide, gave nearly the same number of bands as those that were reduced with dithiothreitol (or beta-mercaptoethanol) and alkylated. However, the mobility of a few proteins was altered by this differential treatment, as visualized by SYPRO-Ruby staining. For example, a protein band(s) migrating at gel position #5 in the alkylated sample is absent in the sample that was both reduced and alkylated. Conversely, proteins observed at gel positions 7 and 8 for the reduced and alkylated sample are not visible in samples only alkylated. Several other differentials occurred between these two treatments; nevertheless, the protein profiles determined for these treatments were similar for the prominent proteins. The use of multiple treatment and separation methods was most useful for low abundant polypeptides as indicated by MASCOT score.

Method 2. Trypsin/HPLC/MS-MS. The trypsin/HPLC/MS-MS method identified 126 virus-encoded proteins, 16 of which were unique to this method. All tryptic or semi-tryptic peptide matches were analyzed using the SPIRE analysis suite (14-17) against PBCV-1 and *C. variabilis* genome databases. Restricting the matches to tryptic only peptides did not decrease the false positive rate, so full semi-tryptic searching was employed. The false positive rate was estimated from searches of a decoy database

of the *Ostreococcus tauri* proteome. The false positive rate was computed to be 0.42%, so one of the 126 proteins identified in this group of experiments might be a false positive. All the proteins identified had a confidence level of 'high' or 'very high' in at least one of the ten analyses in this group and were considered to be in the virion.

Of the ten analyses performed with this method, 6 proteins were detected in only one analysis. One of the proteins was found in 2 analyses, one in 3 analyses, 4 in four analyses, 21 in 5 analyses, 2 in 6 analyses, 2 in 9 analyses, and 89 in all 10 analyses. The number of analyses in which a protein is observed, can be influenced by either variability inherent in mass spectrometry based proteomics experiments, variability in expression, stability of the proteins or false positive results.

Proteome is lower (L) strand and right hand side (R) biased. The genes predicted to encode proteins in the PBCV-1 genome are biased to the right side (262 of 416) relative to the mid-point of the genome; this is also reflected in the number of gene products in the proteome (81 CDSs from right side, 67 CDSs from left side) (Fig. 3). In addition, there is a bias to the reverse strand (L) for the right half of the genome in both the total predicted proteins (159 of the 416, Fig. 3A) and the virion proteome (48 of 148, Fig. 3B). This bias is consistent with certain viable PBCV-1 spontaneous large deletion mutants where up to 40 kbp of the left side of the genome can be deleted (24, 54), and these are recapitulated in the chlorovirus CVK2 (6). The right side L strand virion-coding genes have a mean G+C content of 22%; whereas, the overall G+C content of the genome is 40% and mean G+C content of all the coding genes is 31%. These observations suggest the left side of the genome has less selection pressure relative to the right side for the essential functions of virion assembly and maturation, The right side L

strand is relatively dense with virion-associated genes (38% of the total) with atypical nucleotide composition; whereas, the corresponding left side of the genome is relatively sparse (14%) with regards to virion proteins.

Proteome is skewed to small basic proteins. The PBCV-1 proteome has proteins ranging in molecular weights from 4.9 to 143 kDa and in isoelectric points from 3.6 to 13.0, assuming no post-translational modifications (Fig. 4). Quantitatively, the proteome is dominated by the MCP, centrally located in these distributions. Qualitatively, the proteome is skewed to basic (~75%) and relatively small proteins, approximately 50% are less than 20 kDa, and 63% of the proteins have molecular weights less than 50 kDa and pI values greater than 7.0. This skewing to the more basic side is interesting because the electrostatic charge of the 6 x 10^5 phosphate moieties in the virus genome are probably neutralized by basic proteins (59). However, this prediction must be evaluated further because the stoichiometry of the virion proteins is uncertain. Additionally, how these relate to the chlorovirus CVK2 proteins with DNA binding and protein kinase activities needs to be clarified (60).

Two-dimensional gel analyses using isoelectric focusing versus mass separations support the skewing to basic and small proteins, suggesting that the majority of these proteins are not post-translationally modified in such a way that causes significant deviations of the predicted charge-mass migration (results not shown). However, we never obtained good resolution of the proteins using 2-D gels, even though many protocols were tried, because the MCP dominated the gel.

Membrane proteins. The virion proteins were evaluated for potential transmembrane domains with three independent methods (20, 30, 50); these results

suggest that at least 26% of the proteome may be associated with a membrane structure (Table 1), presumably the internal membrane of the virion. Two-thirds of the CDSs with predicted transmembrane domains (3 out of 3 programs used) were detected by both proteomic methods. The remaining 1/3 of the CDSs were detected equally with Method 1 biased to somewhat larger (mean MW = 23.8 kDa) and more basic proteins (mean pI = 9.2), whereas Method 2 was biased to smaller (mean MW = 10.3 kDa) and less basic proteins (mean pI = 7.8).

The origin of the PBCV-1 internal membrane is unknown. If all, or at least most, of the PBCV-1 internal membrane contains virus-encoded proteins and no host-encoded proteins, it would suggest extensive modification of the host membrane to form the virus membrane.

PBCV-basic adaptor domain containing proteins. Eight PBCV-1 CDSs have at least one copy of a small, highly positively charged C-terminal domain, referred to as the PBCV-basic adaptor domain (19): A092/093L, A176L, A205R, A278L, A282L, A436L, A571R and A676R. All of these CDS were detected in the virion (Table 1). These proteins range in size from 6.9 - 69 kDa, but their pI values are very basic, 10.6 - 13.0. Five of these proteins contain a single copy of the basic adaptor domain; however A092/093L and A278L have 2 copies, and A282L has 3 copies. A278L and A282L are S/T protein kinases (51). The A676R protein contains both the PBCV-basic adaptor domain and a 2-cysteine domain (Pfam 08793), which is a virus-specific domain fused to OUT/A20-like peptidases and S/T protein kinases and is suggested to function as a targeting device for specific substrates (19). The PBCV-basic adaptor domain is only found in the chloroviruses, and A176L is only found in PBCV-1. The function of the

PBCV-basic adaptor domain is unknown.

MCP paralogs. The initial understanding of the architectural makeup of the PBCV-1 virion was a simple quasi-icosahedral particle consisting of a single MCP (Vp54) (64). This picture has evolved to the present 8.5 Å resolution complex particle with several surface features, including a unique vertex with a spike structure and fiber-like structures associated with some capsomers in the trisymmetrons (5, 66). Genome sequencing revealed genes encoding 6 additional capsid-like proteins (26). Previously these paralogs were not considered relevant because at least two of them (genes *a010r* and *a011l*) could be deleted from the genome without loss of virion formation (24). However, the proteome presented here indicates that all of the capsid-like proteins are present in the virion (Table 1) and they fall into 5 paralog classes (Fig. 5A). Each of these proteins contain 2 conserved domains [D1 (green) and D2 (red)] (Fig. 5B) consistent with the Vp54 structure (Fig. 5C). The relative abundance of the proteins, as estimated by their emPAI value, ranged from 1 (A384dL and A383R) to 13 (A430L and A011L). These abundance ratios support the hypothesis that the architecture of-the PBCV-1 virion is composed of a complex mixture of capsids and that the capsomers are composed of heteromeric proteins with a conserved structure. Additionally, the 2 minor capsid-like proteins, A383R and A384dL, contain an additional domain that is similar to the chitin binding peritrophin-A domain (Pfam 01607.17) (SI Table S1) and may contribute to the attachment of the virion to the algal cell surface. The relative abundance of these proteins is consistent with the frequency of fiber structures found in each trisymmetron, but the composition of these structures is unknown.

The estimated relative abundances of virion proteins were determined using the emPAI method (18) for the Method 1 data set. The distribution of the capsid proteins suggests a more complex assembly of PBCV-1 capsids than was previously assumed for a single MCP (Vp54) responsible for the particle architecture. We assume the MCP (A430L) is present in 1440 copies per virion for these calculations and other protein abundances were estimated from this value (Fig. 5B). The data indicate there are two capsid proteins of relatively high abundance (A430L and A011L), two capsid proteins were present at approximately one-half the abundance of these (A010R and A558L), one capsid protein present at one-third abundance (A622L), and two capsid proteins were present in relatively low abundance (A383R and A384dL). Assuming these ratios, icosahedral symmetry, and the fact that the virion is composed of 1680 capsids (64), each of the triangular facets of the icosahedron would contain seven proteins in ratios of 72:72:36:36:24:1:1. Recent structural analysis of PBCV-1 at 8.5 Å resolution indicates the capsomer volumes are more varied than previously thought (66), but how these capsids are arranged is not known. The trimeric capsomers may be homomeric (as previously thought), or possibly heteromeric utilizing the conserved beta-barrel domains as binding surfaces. This higher complexity of virion structure is consistent with several other large DNA viruses where multiple capsid proteins have been detected; herpes viruses have 4 to 7 capsid proteins (22, 33) and mimivirus has at least 5 capsid proteins (37). The emPAI method was used to estimate abundances of intracellular mature virion proteins of vaccinia virus (7) indicating a dynamic range of 1 to 1000 with certain core proteins being most abundant (i. e., A4L, A10L, F17R and A3L), as well as one with low abundance (i. e., E11L).

PBCV-1 proteome functionalities. The 148 virion proteins were grouped into 11 functional/structural categories (SI Fig. S3A) and compared to the distribution of CDSs of the overall genome (SI Fig. S3B). The majority (72%) of virion proteins are in the unknown function category. However, several functions are inferred by sequence similarity analyses and 13 of the 148 proteins have demonstrated functions that include DNA binding, cell signaling via phosphorylation, DNA degradation, virus structure, cell attachment, and polyamine biosynthesis such as homospermidine synthase. Among the identified CDSs are the restriction endonucleases R.CviAII (A252R) and R.CviAI (A579L) thought to be responsible for host DNA degradation early in the infection cycle (3).

Virion morphogenesis is one of the last events in the PBCV-1 replication cycle and it is reasonable that virion proteins are synthesized during the late phase. Most of the proteome (87%) is from genes expressed either late or early-late (65); however, the time of expression has not been determined for 23 new CDSs discovered with the resequence and annotation (SI Fig. S2). Eleven proteins are from genes transcribed in the early phase of replication: 7 of these proteins were detected by a single proteomic method with a relatively low number of unique peptides detected. Therefore, these 7 proteins require further verification. Three of these early proteins, A171R, A440L and A443R have unknown functions. The A456L protein has two conserved domains, a D5 N superfamily domain found in certain viral DNA primases (PfamA: PF08706.4) and a phage/plasmid primase P4 family C-terminal domain with predicted ATPase activity. The A548L protein has two conserved P-loop NTPase domains that are associated with DEXDc-, DEAD- and DEAH-box proteins, including the hepatitis C virus NS3 helicases (PfamA:

PF00176.16). Thus, these proteins might contribute to early transcriptional events that occur within minutes of infection.

PBCV-1 packaged host protein. The PBCV-1 proteome contains one protein (101 amino acids) derived from the host (GenBank: EFN53917.1; 4); the protein was detected by both proteomic methods. This protein is most similar to a fungal 93 amino acid *Naumovozyma dairenensis* CBS 421 nucleosome binding protein (NCBI reference sequence: XP_003667927.1) and similar to the HMGB-UBF_HMG-box, class II and III members of the HMG-box superfamily of DNA-binding proteins. It has no similarity to any PBCV-1 encoded protein. HMG-box containing proteins bind non-B-type DNA conformations with high affinity (45) and they are involved in regulation of DNA-dependent processes such as transcription, replication and DNA repair, all of which require changing the conformation of chromatin (49). Thus, this host protein may be important in initiating PBCV-1 gene expression, which occurs within minutes of infection (65). At least two other large DNA viruses contain chromosomal proteins in the virion. An HMG-box protein (HMG1) and a histone H2B.q protein occur in the Western Reserve strain of vaccinia virus (38) and murine cytomegalovirus virions have a histone H2A protein (22), suggesting large DNA viruses utilize host-derived proteins for DNA binding functions.

Presumed virion proteins that were not detected. A few proteins were expected to be packaged in PBCV-1 that were absent in the proteome analysis. As noted previously, PBCV-1 packages one or more enzymes involved in digesting the host cell wall during infection (29). Annotation of the PBCV-1 genome identified 5 enzymes that might be involved in this process - two chitinases, a chitosanase, a β 1-3 glucanase and a

β & α 1,4 glucuoronic lyase (SI Table S1). Recombinant proteins indicated that all of these enzymes are functional (46, 47) and western blots suggested that one of the chitinases and the chitosanase were in the virion (47). However, none of these five proteins were detected in the proteome analysis. Consequently, the enzyme(s) involved in digesting the host cell wall is unknown.

Circumstantial evidence suggests that PBCV-1 and other chloroviruses package a small virus-encoded K^+ channel protein, named Kcv (12). It has been hypothesized that Kcv is involved in depolarizing the host membrane, which occurs immediately after virus infection. However, Kcv was not detected in this proteome study. On the other hand, at least one putative protein (A201L) with predicted physical/chemical transmembrane properties similar to Kcv was detected in the PBCV-1 virion with proteomic method 1. Thus, this methodology can detect small proteins with transmembrane domains, as in Kcv.

CONCLUSIONS

Re-sequencing and annotation of the 331 kbp chlorovirus PBCV-1 genome revealed that the virus encodes 416 predicted CDSs, using a minimum ORF size of 40 codons, and 11 tRNAs. Proteome analysis of highly purified PBCV-1 virions identified 148 virus-encoded proteins (about 35% of the coding capacity of the virus) and one host protein. Some of these proteins appear to be structural/architectural, whereas others have enzymatic, chromatin modification and signal transduction functions. However, 106 of these proteins have no known function or homologs in the existing gene databases except as orthologs with other chloroviruses, phycodnaviruses and NCLDVs. The genes encoding these proteins are dispersed throughout the virus genome and 84% are

transcribed late or early-late in the infection cycle, which is consistent with virion morphogenesis.

Probably the biggest surprise is that so many virus-encoded proteins were detected in the virion and only one host encoded protein. However, except for the MCP Vp54, we cannot definitively assign a protein(s) to any of the other structural features of the virus, including the additional 6 major capsid-like proteins, the long spike structure, the surface fibers on the trisymmetrons, or the long glue protein homologs of PRD1 and the membrane protein dimer located at the edge of the trisymmetrons and internal membrane (66). These await further structural analyses. Obviously one question is: Are all of these virion-associated proteins essential for creating an infectious virus or are some of them the result of 'sloppy packaging', i. e., fortuitously associated with the particle. This is a difficult question to answer - but it is clear that PBCV-1 morphogenesis is selective in terms of what it incorporates; e.g., the virus packages 2 virus-encoded restriction endonucleases, but not their corresponding DNA methyltransferases. In addition, only one host protein was detected in the virion; no host membrane proteins were detected.

The PBCV-1 capsid protein composition may be somewhat flexible because the genes encoding 2 of the capsid proteins (A010R and A011L) can be deleted (6, 24), yet these deletion mutants are viable. This finding suggests some type of compensation in capsid protein utility. Among large DNA viruses, the number of capsid proteins ranges from 4 to 7 and these homologs are virion-associated (e. g., 22, 33, 37), thus the discovery of 7 putative capsid proteins in the PBCV-1 virion is consistent with this theme yet little is known how these proteins contribute to virion structure or function.

These data are essential for understanding the immediate-early events of infection such as binding, entrisome, macromolecular synthetic shutoff, DNA degradation, viral transcription, etc. (48). PBCV-1 is emblematic of the giant viruses in that they are large, complex, and highly diverse (11). When characterizing these viruses, the traditional meaning of the term "structural protein" (i. e., virion-associated) has lost its meaning. Giant virus virions incorporate both structural/architectural proteins as well as many other proteins with a wide range of functionalities likely directed at the immediate-early events of infection before viral transcription *de novo* is initiated. Why giant viruses have so many genes continues to intrigue virologists, but as more giant viruses are discovered there will be more opportunities to explore this question.

ACKNOWLEDGEMENTS

The UNL mass spectrometry facility is supported in part by NIH Grant P20 RR15635 from the COBRE Program of the National Center for Research Resources, NCI Cancer Center Support Grant P30 CA36727, and the Nebraska Research Initiative. Additional support was from NIH COBRE Grant P20RR015635 (DDD), NIH R0l GM32441 (JLVE), NSF-EPSCoR EPS-1004094 (JLVE), DOE DE-EE0003142 (JLVE), Discovery grant from NSERC (Canada) (CU), National Center for Research Resources 5P20RR016469 and NIGMS 8P20GM103427 (GD).

We thank Jaehyoung Kim and Joe Nietfeldt at the UNL Core for Applied Genomics and Ecology for the PBCV-1 sequencing.

REFERENCES:

1. **Abrescia, N. G., J. J. Cockburn, J. M. Grimes, G. C. Sutton, J. M. Diprose, S. J. Butcher, S. D. Fuller, C. San Martín, R. M. Burnett, D. I. Stuart, D. H. Bamford, and J. K. Bamford.** 2004. Insights into assembly from structural analysis of bacteriophage PRD1. Nature **432:**68-74.
2. **Agarkova, I., D. Dunigan, J. Gurnon, T. Greiner, J. Barres, G. Thiel, and J. Van Etten.** 2008. Chlorovirus-mediated membrane depolarization of *Chlorella* alters secondary active transport of solutes. J. Virol. **82:**12181-12190.
3. **Agarkova, I. V., D.D. Dunigan, J.L. Van Etten.** 2006. Virion-associated restriction endonucleases of chloroviruses. J. Virol. **80:**8114-8123.
4. **Blanc, G., G. Duncan, I. Agarkova, M. Borodovsky, J. Gurnon, A. Kuo, E. Lindquist, S. Lucas, J. Pangilinan, J. Polle, A. Salamov, A. Terry, T. Yamada, D. Dunigan, I. Grigoriev, J.-M. Claverie, and J. L. Van Etten.** 2010. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell **22:**2943-2955.
5. **Cherrier, M. V., V. A. Kostyuchenko, C. Xiao, V. D. Bowman, A. J. Battisti, X. Yan, P. R. Chipman, T. S. Baker, J. L. Van Etten, and M. G. Rossmann.** 2009. An icosahedral algal virus has a complex unique vertex decorated by a spike. Proc Natl Acad Sci USA **106:**11085-11089.
6. **Chuchird, N., K. Nishida, T. Kawasaki, M. Fujie, S. Usami, and T. Yamada.** 2002. A variable region on the chlorovirus CVK2 genome contains five copies of the gene for Vp260, a viral-surface glycoprotein. Virology **295:**289-298.
7. **Chung, C.-S., C.-H. Chen, M.-Y. Ho, C.-Y. Huang, C.-L. Liao, and W.**

Chang. 2006. Vaccinia virus proteome: identification of proteins in vaccinia virus intracellular mature virion particles. J. Virol. **80:**2127-2140.

8. **Cockburn, J., N. Abrescia, J. Grimes, G. Sutton, J. Diprose, J. Benevides, G. J. Thomas, J. Bamford, D. Bamford, and D. Stuart.** 2004. Membrane structure and interactions with protein and DNA in bacteriophage PRD1. Nature **432:**122-125.

9. **Craig, R., and R. C. Beavis.** 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics **20:**1466-1467.

10. **Dereeper, A., V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J. F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. M. Claverie, and O. Gascuel.** 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. **36:**W465-469.

11. **Dunigan, D. D., L. A. Fitzgerald, and J. L. Van Etten.** 2006. Phycodnaviruses: a peek at genetic diversity. Virus Res. **117:**119-132.

12. **Frohns, F., A. Käsmann, D. Kramer, B. Schäfer, M. Mehmel, M. Kang, J.L. Van Etten, S. Gazzarrini, A. Moroni, G. Thiel.** 2006. Potassium ion channels of chlorella viruses cause rapid depolarization of host cells during infection. J. Virol. **80:**2437-2444.

13. **Graves, M. V., and R. H. Meints.** 1992. Characterization of the major capsid protein and cloning of its gene from algal virus PBCV-1. Virology **188:**198-207.

14. **Higdon, R., G. Hather, A. T. Bauman, B. Louie, B. Broomall, S. Fortenly, N. Kolker, G. van Belle, and E. Kolker.** 2009. SPIRE: systematic protein identification and relative expression analysis resource for high-throughput

proteomics, 57th ASMS Conference on Mass Spectrometry, Philadelphia, Pennsylvania, May 31 - June 4, 2009.

15. **Higdon, R., J. Hogan, N. Kolker, G. van Belle, and E. Kolker.** 2007. Experiment-specific estimation of peptide identification probabilities using a randomized database. OMICS **11:**351-356.

16. **Higdon, R., J. Hogan, G. Van Belle, and E. Kolker.** 2005. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. OMICS **9:**364-379.

17. **Higdon, R., and E. Kolker.** 2007. A predictive model for identifying proteins by a single peptide match. Bioinformatics **23:**277-280.

18. **Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann.** 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell Proteomics **4:**1265-1272.

19. **Iyer, L., S. Balaji, E. Koonin, and L. Aravind.** 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. **117** 156-184.

20. **Käll, L., A. Krogh, and E. L. Sonnhammer.** 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. **338:**1027-1036.

21. **Kang, M., D. D. Dunigan, and J. L. Van Etten.** 2005. Chloroviruses: a genus of Phycodnaviridae that infect certain chlorella-like green algae. Mol. Plant Path. **6:**213-224.

22. **Kattenhorn, L., R. Mills, M. Wagner, A. Lomsadze, V. Makeev, M. Borodovsky, H. Ploegh, and B. Kessler.** 2004. Identification of proteins

associated with murine cytomegalovirus virions. J. Virol. **78:**11187-11197.

23. **Kuznetsov, Y. G., J. R. Gurnon, J. L. Van Etten, and A. McPherson.** 2005. Atomic force microscopy investigation of a chlorella virus, PBCV-1. J. Struct. Biol. **149:**256-263.

24. **Landstein, D., D. E. Burbank, J. W. Nietfeldt, and J. L. Van Etten.** 1995. Large deletions in antigenic variants of the chlorella virus PBCV-1. Virology **214:**413-420.

25. **Lane, L.** 1978. A simple method for stabilizing protein-sulfhydryl groups during SDS-gel electrophoresis. Anal. Biochem. **86:**655-664.

26. **Li, Y., Z. Lu, L. Sun, S. Ropp, G. F. Kutish, D. L. Rock, and J. L. Van Etten.** 1997. Analysis of 74 kb of DNA located at the right end of the 330-kb chlorella virus PBCV-1 genome. Virology **237:**360-377.

27. **Mehmel, M., M. Rothermel, T. Meckel, J.L. Van Etten, A. Moroni, G. Thiel.** 2003. Possible function for virus encoded K^+ channel Kcv in the replication of chlorella virus PBCV-1. FEBS Lett. **552:**7-11.

28. **Meints, R. H., D. E. Burbank, V. E. J.L., and L. D.T.** 1988. Properties of the chlorella receptor for the virus PBCV-1. Virology **164:**15-21.

29. **Meints, R. H., K. Lee, D.E. Burbank, J.L. Van Etten.** 1984. Infection of a chlorella-like alga with the virus, PBCV-1: Ultrastructural studies. Virology **138:**341-346.

30. **Möller, S., M. D. R. Croning, and R. Apweiler.** 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics **17:**646-653.

31. **Nandhagopal, N., A. A. Simpson, J. R. Gurnon, X. Yan, T. S. Baker, M. V.**

Graves, J. L. Van Etten, and M. G. Rossmann. 2002. The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. Proc Natl Acad Sci USA **99:**14758-14763.

32. **Neupärtl, M., C. Meyer, I. Woll, F. Frohns, M. Kang, J.L. Van Etten, D. Kramer, B. Hertel, A. Moroni, G. Thiel.** 2008. Chlorella viruses evoke a rapid release of K^+ from host cells during early phase of infection. Virology **372:**340-348.

33. **O'Connor, C. M., and D. H. Kedes.** 2006. Mass spectrometric analyses of purified rhesus monkey rhadinovirus reveal 33 virion-associated proteins. J. Virol. **80:**1574-1583.

34. **Onimatsu, H., K. Suganuma, S. Uenoyama, and T. Yamada.** 2006. C-terminal repetitive motifs in Vp130 present at the unique vertex of the chlorovirus capsid are essential for binding to the host chlorella cell wall. Virology **353:**432-442.

35. **Onimatsu, H., I. Sugimoto, M. Fujie, S. Usami, and T. Yamada.** 2004. Vp130, a chloroviral surface protein that interacts with the host Chlorella cell wall. Virology **319:**71-80.

36. **Que, Q., Y. Li, I. N. Wang, L. C. Lane, W. G. Chaney, and J. L. Van Etten.** 1994. Protein glycosylation and myristylation in chlorella virus PBCV-1 and its antigenic variants. Virology **203:**320-7.

37. **Renesto, P., C. Abergel, P. Decloquement, D. Moinier, S. Assa, H. Ogata, P. Foruquet, J. P. Gorvel, and J. M. Claverie.** 2006. Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J. Virol. **80:**11678-11685.

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

39. **Santoni, V., M. Molloy, and T. Rabilloud.** 2000. Membrane proteins and proteomics: Un amour impossible? Electrophoresis **21:**1054-1070.

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

41. **Short, C. M., O. Rusanova, and S. M. Short.** 2011. Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. ISME J. **5:**810-821.

42. **Skrdla, M. P., D. E. Burbank, Y. Xia, R. H. Meints, and J. L. Van Etten.** 1984. Structural proteins and lipids in a virus, PBCV-1, which replicates in a chlorella-like alga. Virology **135:**308-315.

43. **Song, W. J., Q. W. Qin, J. Qiu, C. H. Huang, F. Wang, and C. L. Hew.** 2004. Functional genomics analysis of Singapore grouper iridovirus: complete sequence determination and proteomic analysis. J. Virol. **78:**12576-12590.

44. **Stothard, P., and D. S. Wishart.** 2005. Circular genome visualization and exploration using CGView. Bioinformatics **21:**537-539.

45. **Stros, M., D. Launholt, and K. D. Grasser.** 2007. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell. Mol. Life Sci. **64:**2590-2606.

46. **Sugimoto, I., H. Onimatsu, M. Fujie, S. Usami, and T. Yamada.** 2004. vAL-1, a novel polysaccharide lyase encoded by chlorovirus CVK2. FEBS Lett. **559:**51-

56.

47. **Sun, L., B. Adams, J. Gurnon, Y. Ye, and J. L. Van Etten.** 1999. Characterization of two chitinase genes and one chitosanase gene encoded by chlorella virus PBCV-1. Virology **263:**376-387.

48. **Thiel, G., A. Moroni, D. Dunigan, and J. L. Van Etten.** 2010. Initial events associated with virus PBCV-1 infection of *Chlorella* NC64A. Prog. Bo. **71:**169-183.

49. **Thomas, J. O.** 2001. HMG1 and 2: architectural DNA-binding proteins. Biochem. Soc. Trans. **29:**395-401.

50. **Tusnády, G. E., and I. Simon.** 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics **17:**849-850.

51. **Valbuzzi, P.** 2005. Serine/threonine kinases encoded by PBCV-1 virus: characterization and possible role in the phosphorylation of the K+ channel Kcv. Ph. D. Dissertation. Universita Degli Studi di Milano, Italy.

52. **Van Etten, J. L., D. E. Burbank, D. Kuczmarski, and R. H. Meints.** 1983. Virus infection of culturable chlorella-like algae and development of a plaque assay. Science **219:**994-996.

53. **Van Etten, J. L., D.E. Burbank, Y. Xia, R.H. Meints.** 1983. Growth cycle of a virus, PBCV- 1, that infects chlorella-like algae. Virology **126:**117-125.

54. **Van Etten, J. L., and D. D. Dunigan.** 2012. Chloroviruses: not your every day plant virus. Trends Plant Sci. **17:**1-8.

55. **Van Etten, J. L., M. V. Graves, D. G. Muller, W. Boland, and N. Delaroque.** 2002. Phycodnaviridae-large DNA algal viruses. Arch. Virol. **147:**1479-1516.

56. **Van Etten, J. L., L. C. Lane, and R. H. Meints.** 1991. Viruses and viruslike particles of eukaryotic algae. Microbiol. Rev. **55:**586-620.

57. **Van Etten, J. L., and R. H. Meints.** 1999. Giant viruses infecting algae. Annu. Rev. Microbiol. **53:**447-494.

58. **Van Etten, J. L., C. H. Van Etten, J. K. Johnson, and D. E. Burbank.** 1985. A survey for viruses from fresh water that infect a eucaryotic chlorella-like green alga. Appl. Environ. Microbiol. **49:**1326-1328.

59. **Wulfmeyer, T., C. Polzer, G. Hiepler, K. Hamacher, R. Shoeman, D. D. Dunigan, J. L. Van Etten, M. Lolicato, A. Moroni, G. Thiel, and T. Meckel.** 2012. Structural organization of DNA in chlorella viruses. PLoS One **7:**e30133.

60. **Yamada, T., S. Furukawa, T. Hamazaki, and P. Songsri.** 1996. Characterization of DNA-binding proteins and protein kinase activities in chlorella virus CVK2. Virology **219:**395-406.

61. **Yamada, T., T. Higashiyama, and T. Fukuda.** 1991. Screening of natural waters for viruses which infect chlorella cells. Appl. Environ. Microbiol. **57:**3433-3437.

62. **Yamada, T., H. Onimatsu, and J. L. Van Etten.** 2006. Chlorella viruses, p. 293-366. *In* K. Maramorosch and A. J. Shatkin (ed.), Adv. Virus Res., vol. 66. Elsevier Inc.

63. **Yan, X., V. Bowman, N. H. Olson, J. R. Gurnon, J. L. Van Etten, M. G. Rossmann, and T. S. Baker.** 2005. The structure of a T=169d algal virus, PBCV-1, at 15 Â resolution. Microsc. Microanal. **11:**1056-1057.

64. **Yan, X., N. H. Olson, J. L. Van Etten, M. Bergoin, M. G. Rossmann, and T.**

S. Baker. 2000. Structure and assembly of large lipid-containing dsDNA viruses. Nat. Struct. Biol. **7:**101-103.

65. **Yanai-Balser, G. M., G. A. Duncan, J. D. Eudy, D. Wang, X. Li, I. V. Agarkova, D. D. Dunigan, and J. L. Van Etten.** 2010. Microarray analysis of chlorella virus PBCV-1 transcription. J.Virol. **84:**532-542.

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

67. **Zhang, Y., I. Calin-Jageman, J. R. Gurnon, T. J. Choi, B. Adams, A. W. Nicholson, and J. L. Van Etten.** 2003. Characterization of a chlorella virus PBCV-1 encoded ribonuclease III. Virology **317:**73-83.

68. **Zhang, Y., P. Strasser, R. Grabherr, and J. L. Van Etten.** 1994. Hairpin loop structure at the termini of the chlorella virus PBCV-1 genome. Virology **202:**1079-1082.

Figure legends

Figure 1. Proteomic methodologies for PBCV-1 virions.

Figure 2. SDS-PAGE protein separation and virion proteome mapped onto the PBCV-1 genome. The PBCV-1 genome was re-sequenced, assembled and annotated to correct existing sequence errors. The 416 predicted CDSs are represented as grey arrows running both clockwise and counter-clockwise along the genome (panel B). Note: the diagram is circular, but there is a break at the 12 o'clock position because the viral genome is a linear molecule with terminal inverted repeats and closed hairpin ends. The terminal sequences (inverted repeats and hairpin ends) were found to be identical to those reported previously (68). The polycistronic gene encoding 11 tRNAs is presented in red (see 6 o'clock). The 148 proteins of the virion proteome were determined using two independent mass spectrometry-based methods (see Materials and Methods). The results of each method are shown, proteins determined uniquely by method 1 are presented in magenta, proteins determined uniquely by method 2 are presented in blue, proteins determined by both methods 1 and 2 are presented in brown. The map was developed from the CGView software (44). Panel A shows the distribution of virion proteins with SDS polyacrylamide gel separation. The numbers to the left indicate the gel fragment that was analyzed.

Figure 3. Expression stage distribution of PBCV-1 CDSs, a quartile analysis. The number of all coding CDSs (panel A) expressed either during the early (blue), early-late (red), late (green), or not determined (nd, purple) is shown as a function of the genome map position. The genome map is divided into four regions, both the direct ("R" genes)

834 and reverse ("L" genes) on each half of the genome (left half gene numbers: 001 – 327;
835 right half gene numbers: 328 – 692). Panel B shows the distribution of virion-associated
836 CDSs with respect to expression stage and genome position.

837 **Figure 4**. Mass versus pI distribution of PBCV-1 virion CDSs identified by two
838 independent proteomic methods. The virion proteins are displayed as a function of their
839 intrinsic molecular weight and isoelectric point. The results of each method are shown,
840 proteins determined uniquely using method 1 are presented in magenta, proteins
841 determined uniquely using method 2 are presented in blue, proteins determined using
842 both methods 1 and 2 are presented in brown. Note that Method 2 was especially useful
843 for discovering a set of low molecular weight proteins that were not detected with
844 Method 1.

845 **Figure 5**. Capsid protein paralog classes and relative abundances in PBCV-1. The seven
846 capsid-like proteins detected in the PBCV-1 virion were evaluated against a dataset of
847 chloroviruses, including PBCV-1 (RefSeq NC_000852.5), NY-2A (RefSeq
848 NC_009898.1), AR158 (RefSeq NC_009899.1), MT325 (GenBank DQ491001.1), FR483
849 (RefSeq NC_008603.1), and ATCV-1 (RefSeq NC_008724.1). These 7 proteins had
850 homologs in each of the viruses that separate into 5 distinct paralog classes (I – V) as
851 shown in the neighbor joining tree (panel A) (see SI Table S3 for CDS accession
852 numbers). The sequence for PBCV-1 A384dL, a member of paralog class V which is
853 distantly related, was used as the out-group to root the phylogenetic analysis using the
854 website www.phylogeny.fr (10). Muscle was used to align the sequences. Bootstrap
855 analysis was used to construct the tree. Similar tree topologies were produced by

maximum likelihood and maximum parsimony analyses. The values on the branches are the percentage of bootstrap support (200 replicates). Only bootstrap values >50% are shown. The distance bar represents 0.2 amino acid substitution per site. Panel B presents the PBCV-1 capsid proteins grouped into 5 paralog classes within their two conserved domains. The D1 domain (green, column A) and the D2 (red, column D) NCLDV superfamily capsid domain were previously determined by structure analysis of the Vp54 MCP (31) (panel C). The relative abundances as determined with the emPAI method for the Method 1 data are listed to the right of the table, as well as the hypothetical estimated copies per virion of each capsid protein. Note, the two proteins of relatively lower abundance contain chitin binding peritrophin-A conserved domains (columns C and E).

866 Table 1. PBCV-1 virion proteome

Protein (CDS)	Da	pI	Expression stage	Function or putative function	Proteomic method	TM prediction[a] T	H	P
A010R	44998	5.2	Late	Capsid protein; PfamA: PF4451.5 [1.9e-50]	1&2	0	0	0
A011L	45076	5.4	Late	Capsid protein; PfamA: PF4451.5 [2.9e-61]	1&2	0	0	0
A014R	141382	6.3	Late	Unknown protein	1&2	0	0	0
A018L	137639	4.9	Late	Unknown protein; PfamA: PF06598.4 [Chlorovirus glycoprotein repeat] [1.2e-11]	1	0	0	0
A025/027/029L	140095	4.4	Late	Unknown protein	1&2	0	0	0
A034R	35163	10.4	Late	Protein kinase; PfamA: PF00069.18 [Protein kinase domain] [1.4e-07]	1&2	0	1	0
A035L	65606	8.9	Late	Unknown protein	1&2	0	1	0
A041R	44315	10.8	Late	Unknown protein	1&2	0	1	0
A051L	22804	8.6	Late	Unknown protein	1&2	1	2	1
A085R	27812	7.8	Late	Prolyl 4-hydroxylase; PfamA: PF03171.13 [2OG-Fe(II) oxygenase superfamily] [3.5e-11]	1&2	1	1	1
A092/093L	49577	10.7	Early-Late	Unknown protein; PfamA: PF08789.3 [PBCV-specific basic adaptor domain] [1.2e-15]	1&2	0	0	0
A121R	12486	10.8	Early-Late	Unknown protein	1&2	0	0	0
A122/123cL	4912	10.1	N/A	Unknown protein	1	0	0	0
A122/123R	137880	5.0	Late	COG5295 [Autotransporter adhesin] [4e-12]; PfamA: PF06598.4 [Chlorovirus glycoprotein repeat] [3.6e-11] / PF11962.1 [Domain of unknown function (DUF3476)] [8.2e-66]	1	0	31	0
A127R	27126	10.1	Late	Unknown protein	1&2	0	0	0
A136R	16367	11.5	N/A	Unknown protein	1&2	0	0	0
A137R	8777	10.9	Early	Unknown protein	1	0	0	0

A139L	17701	8.4	Late	Unknown protein	1&2	2	2	2
A140/145R	120898	11.0	Early-Late	Unknown protein	1&2	0	1	0
A157L	12328	3.9	Early-Late	Unknown protein	2	1	1	1
A164aR	7094	5.8	N/A	Unknown protein	2	1	0	0
A165aL	19024	10.1	N/A	Unknown protein	1&2	0	0	0
A168R	18317	4.6	Late	Unknown protein	1&2	1	1	1
A171R	42413	10.2	Early	Unknown protein	1&2	0	0	0
A172aL	6053	9.8	N/A	Unknown protein	1	1	1	0
A173L	31933	8.2	Early	COG1752 [Predicted esterase of the alpha-beta hydrolase superfamily] [2e-06]; PfamA: PF01734.15 [Patatin-like phospholipase] [4.2e-27]	1	0	2	0
A174L	7453	12.2	N/A	Unknown protein	2	0	0	0
A176L	9167	11.3	N/A	Unknown protein; PfamA: PF08789.3 [PBCV-specific basic adaptor domain] [9e-12]	1&2	0	0	0
A188aR	17326	10.0	N/A	COG0417 [DNA polymerase elongation subunit (family B)] [3e-07]; PfamA: PF00136.14 [DNA polymerase family B] [6.5e-17]	1	0	0	0
A189/192R	143575	11.4	Late	Unknown protein	1&2	0	0	0
A196L	17456	8.4	Late	Unknown protein	2	3	3	1
A201aL	6787	8.8	N/A	Unknown protein	1	0	0	0
A201L	10005	10.7	Early-Late	Unknown protein	1	2	2	2
A202L	12232	5.0	Early-Late	Unknown protein	2	0	0	0
A203R	24011	6.0	Late	Unknown protein	1&2	1	2	0
A205R	22452	12.1	Late	Unknown protein; PfamA: PF08789.3 [PBCV-specific basic adaptor domain] [4.2e-16]	1&2	0	0	0
A213L	16483	4.5	Early-Late	Unknown protein	1&2	1	1	1
A217L	45248	9.9	Early-Late	Unknown protein	1&2	0	0	1
A219/222/226R	77797	7.0	Early	COG1215 [Glycosyltransferases probably involved in cell wall biogenesis] [4e-06]; Swissprot: P58932 [RecName: FullCellulose synthase catalytic subunit (UDP-forming)]	1	9	8	10

				[6e-07]				
A227L	15689	10.0	Late	Unknown protein	1&2	0	0	0
A230R	22055	8.4	Late	Unknown protein	1&2	4	4	4
A231L	43644	9.9	Early-Late	Unknown protein	1&2	1	0	0
A237R	58565	9.5	Late	Homospermidine synthase	1&2	0	0	0
A245R	19748	9.3	Late	Cu/Zn superoxide dismutase	1&2	1	1	0
A246R	12017	11.5	Late	Unknown protein	1&2	0	0	0
A252R	39856	10.3	Early	R.CviAII restriction endonuclease	1&2	0	0	0
A255R	17300	5.1	N/A	Unknown protein	1	0	0	0
A256/257L	96729	7.2	Early-Late	Unknown protein	1	0	0	0
A260aR	7742	11.9	N/A	Unknown protein	1	0	0	0
A262/263L	29470	9.6	N/A	Unknown protein	1&2	2	3	2
A271L	31114	7.1	Early-Late	COG2267 [Lysophospholipase] [1e-07]	1	0	3	0
A273L	15713	9.9	Late	PF03713.6 [Domain of unknown function (DUF305)] [6.8e-13]	1	3	3	3
A278L	69231	10.8	Late	Protein kinase; PfamA: PF00069.18 [Protein kinase domain] [1.2e-07] / PF08789.3 [PBCV-specific basic adaptor domain] [7.5e-10]	1&2	0	1	0
A282L	63371	10.8	Late	Protein kinase; PfamA: PF00069.18 [Protein kinase domain] [1.2e-07] / PF08789.3 [PBCV-specific basic adaptor domain] [1.3e-17]	1&2	0	1	0
A284L	30766	9.2	Early-Late	Amindase	1&2	0	0	0
A286R	43042	9.6	Late	Unknown protein	1&2	0	0	0
A287R	31349	9.4	Early-Late	PfamA: PF01541.17 [GIY-YIG catalytic domain] [4.2e-11] / PF07453.6 [NUMOD1 domain] [8.6e-11]	1	0	0	0
A295L	35626	7.9	Early-Late	Fucose synthetase; Swissprot: Q9LMU0 [RecName: FullPutative GDP-L-fucose synthase 2 AltName: FullGDP-4-keto-6-deoxy-D-mannose-3 5-epimerase-4-reductase 2 ShortAtGER2] [1e-100]	1	0	0	0

A296R	17393	12.2	Late	Unknown protein	1&2	0	1	1
A304R	9490	5.8	Late	Unknown protein	1	0	0	0
A305L	22910	10.7	Late	Protein phosphatase; Swissprot: Q9BY84 [RecName: FullDual specificity protein phosphatase 16 AltName: FullMitogen-activated protein kinase phosphatase 7 ShortMAP kinase phosphatase 7 ShortMKP-7] [7e-12]	1&2	0	0	0
A310L	18268	8.5	Late	Unknown protein	1&2	0	0	0
A314R	9114	6.7	Late	Unknown protein	1&2	1	1	1
A316R	48779	10.7	Late	Unknown protein	1&2	0	1	0
A320R	15685	10.5	Late	Unknown protein	1&2	1	1	1
A321R	12830	8.8	Late	Unknown protein	1	2	2	2
A322L	20039	5.0	Late	Unknown protein	1&2	1	1	1
A339L	7372	11.1	Early-Late	Unknown protein	1	0	0	0
A342L	63813	9.2	Early-Late	Unknown protein	1&2	1	1	1
A349L	21077	10.0	Early-Late	Unknown protein	1&2	0	1	0
A350R	14676	9.7	N/A	PfamA: PF12239.1 [Protein of unknown function (DUF3605)] [4.4e-23]	2	0	0	0
A352L	23310	3.6	Late	Swissprot: Q5UQF7 [RecName: FullUncharacterized protein R489 Flags: Precursor] [1e-05]	1&2	0	1	1
A356R	12512	10.5	N/A	Unknown protein	1	0	0	0
A363R	128448	10.9	Early	Swissprot: P0C9B2 [RecName: FullPutative ATP-dependent RNA helicase Q706L] [2e-06]	1&2	0	2	0
A375R	19085	9.4	Early-Late	Unknown protein	1&2	2	2	2
A378L	29219	9.4	Late	Unknown protein	1&2	1	1	0
A383R	52511	5.2	Late	Capsid protein; Pfam: PF04451.5 [Large eukaryotic DNA virus major capsid protein] [1.6e-25]	1&2	0	0	0
A384bL	6809	9.0	N/A	Unknown protein	2	1	1	1
A384dL	69009	8.0	Early-Late	Capsid protein; PfamA: PF01607.17 [Chitin binding Peritrophin-A domain] [2.4e-07] / PF04451.5 [Large eukaryotic DNA virus	1&2	1	2	1

				major capsid protein] [2e-11]				
A398L	12987	9.9	Late	Unknown protein	1&2	2	3	3
A400R	13634	9.5	Early-Late	Unknown protein	2	0	0	0
A405R	53502	10.3	Late	Unknown protein	1&2	1	2	1
A407L	23382	8.9	Late	Unknown protein	1&2	1	2	2
A413L	26998	9.5	Late	Unknown protein	1&2	2	2	2
A414R	10612	10.8	Late	Unknown protein	1&2	2	2	2
A420L	7918	6.4	Late	Unknown protein	2	1	1	1
A421R	11056	10.1	Late	Unknown protein	1&2	1	1	1
A423R	18458	6.5	Late	Unknown protein	2	0	1	0
A430L	48165	7.5	Late	Major capsid protein	1&2	0	0	0
A436L	6932	13.0	N/A	Unknown protein; Pfam: PF08789.3 [PBCV-specific basic adaptor domain] [1.5e-16]	1	0	0	0
A437L	10876	11.0	Late	PfamA: PF05854.4 [Non-histone chromosomal protein MC1] [5.9e-07]	1&2	0	1	0
A438L	8988	10.7	Early-Late	Glutaredoxin	2	0	0	0
A440L	10112	11.1	Early	Unknown protein	1&2	0	0	0
A443R	34961	5.3	Early	Unknown protein	1	0	0	0
A448L	12369	10.4	Late	Protein disulphide isomerase with heme binding site	1&2	0	0	0
A454L	31194	4.7	Early-Late	Unknown protein	1&2	1	1	0
A456L	75235	5.5	Early	COG3378 [Predicted ATPase] [3e-06]; PfamA: PF08706.4 [D5 N terminal like] [3.9e-09]	1	0	0	0
A465R	13528	10.2	Early-Late	COG5054 [Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins] [4e-06]; PfamA: PF04777.6 [Erv1 / Alr family] [3.5e-22]	1&2	0	0	0
A476R	37393	4.4	Early-Late	Swissprot: Q6Y657 [RecName: FullPutative ribonucleoside-diphosphate reductase small chain B AltName: FullRibonucleotide reductase small subunit B AltName: FullRibonucleoside-diphosphate reductase	1	0	0	1

R2B subunit] [1e-113]

A480L	9838	10.0	Late	Unknown protein	1&2	2	2	2
A484L	18604	9.6	Early-Late	Unknown protein	1&2	0	0	0
A488R	34631	5.0	Late	Swissprot: Q5UQL4 [RecName: FullUncharacterized protein L417] [2e-09]	1&2	0	3	0
A497R	15378	10.4	Late	Unknown protein	1&2	2	2	1
A500L	38463	5.0	N/A	Unknown protein	1&2	1	2	1
A502L	11069	9.4	Late	Unknown protein	2	1	1	1
A520L	11674	10.7	Late	Unknown protein	2	0	0	0
A521aL	22578	6.3	N/A	Swissprot: O55742 [RecName: FullUncharacterized protein 136R] [2e-07]	1&2	0	0	0
A521L	23738	11.4	Early-Late	Unknown protein	1&2	0	0	0
A523R	19096	9.6	Late	Unknown protein	1&2	0	0	0
A526R	16434	9.3	Late	Unknown protein	1&2	0	1	0
A527R	11605	10.7	Late	Unknown protein	1&2	0	0	0
A531L	7670	7.5	Late	Unknown protein	2	1	1	1
A532aL	5479	4.5	N/A	Unknown protein	2	1	1	1
A532L	8698	9.7	Late	Unknown protein	1&2	1	1	1
A533R	40132	3.8	Early-Late	Unknown protein	1&2	0	0	0
A534R	11783	9.7	N/A	Unknown protein	1&2	0	0	0
A535L	8210	4.7	Early-Late	Unknown protein	1&2	0	0	0
A536L	8485	10.0	Early-Late	Unknown protein	1&2	1	1	0
A540L	127197	6.2	Late	Unknown protein	1	0	0	0
A548L	57432	9.5	Early	PfamA: PF00176.16 [SNF2 family N-terminal domain] [6.7e-34] / PF00271.24 [Helicase conserved C-terminal domain] [1.5e-10]	1	0	0	0
A558L	45547	5.1	Early-Late	Capsid protein; PfamA: PF04451.5 [Large eukaryotic DNA virus major capsid protein] [6.6e-60]	1&2	0	0	0
A559L	24034	10.2	Late	Unknown protein	1&2	1	1	0

A561L	71004	9.9	Late	Unknown protein	1&2	1	2	1
A565R	73169	7.3	Early-Late	Unknown protein	1&2	1	1	1
A567L	17418	10.1	Early-Late	Unknown protein	1	0	0	0
A571R	12972	12.0	Late	Pfam hit: PF08789.3 [PBCV-specific basic adaptor domain] [5.7e-17]; Refseq best hit: YP_001426112 [hypothetical protein FR483_N480R (Paramecium bursaria Chlorella virus FR483)] [3e-39]	1	0	0	0
A572R	20606	7.1	Late	Unknown protein	1&2	0	0	0
A577L	15442	11.0	Late	Unknown protein	1&2	0	0	0
A579L	27445	10.1	Late	R.CviAI restriction endonuclease	1&2	0	0	0
A586R	8567	11.8	N/A	Unknown protein	1	0	0	0
A598L	41558	6.9	Early-Late	COG0076 [Glutamate decarboxylase and related PLP-dependent proteins] [5e-06]; PfamA: PF00282.12 [Pyridoxal-dependent decarboxylase conserved domain] [1.1e-17]	1	0	0	0
A605L	17769	10.9	Early-Late	Unknown protein	1&2	1	1	1
A612L	13587	8.7	Late	Histone H3K27 methylase	2	0	0	0
A614L	64733	11.2	Late	Protein kinase; PfamA: PF00069.18 [Protein kinase domain] [5.6e-11]	1&2	0	0	0
A617R	37586	9.9	Early-Late	Swissprot: Q5UQJ6 [RecName: FullPutative serine/threonine-protein kinase R400] [7e-12]	1	0	0	0
A621L	12935	9.5	Late	Unknown protein	1	2	2	2
A622L	58097	5.7	Late	Capsid protein; PfamA: PF04451.5 [Large eukaryotic DNA virus major capsid protein] [1.7e-66]	1&2	0	0	0
A624R	13570	9.3	Late	Unknown protein; PfamA: PF09945.2 [Predicted membrane protein (DUF2177)] [3.4e-26]	1	3	4	3
A625R	49945	10.7	Late	COG0675 [Transposase and inactivated derivatives] [1e-06]; PfamA: PF12323.1 [Helix-turn-helix domain] [1.4e-06] / PF07282.4 [Putative transposase DNA-binding domain] [6.7e-18]	1	0	0	0
A627R	49629	11.1	Late	Unknown protein	1&2	1	3	0

A629R	86292	7.5	Early-Late	PfamA: PF03477.9 [ATP cone domain] [8.5e-15] / PF00317.14 [Ribonucleotide reductase all-alpha domain] [7.9e-19] / PF02867.8 [Ribonucleotide reductase barrel domain] [2e-194]	1	0	0	0
A631L	10392	9.9	N/A	Unknown protein	1	0	0	0
A643R	53097	11.3	Late	Unknown protein	1&2	0	0	0
A644R	19207	6.0	Late	Unknown protein	1&2	0	0	0
A655L	12002	11.4	N/A	Unknown protein	1	0	1	0
A676R	42432	10.6	Late	Unknown protein; PfamA: PF08789.3 [PBCV-specific basic adaptor domain] [1.9e-17] / PF08793.3 [2-cysteine adaptor domain] [1.8e-15]	1&2	0	0	0
A678R	41287	10.3	Late	Unknown protein	1&2	0	3	0
A686L	18316	6.9	Early	Unknown protein	1	0	1	0

a – Transmembrane regions of the protein were predicted by TMHMM [T] (30), HMMTOP [H] (50), and Phobius [P] (20) methods. For all the method default parameters were used for the prediction. The number shown in the table is the number of helixes predicted by the method.

Virus purification:
Differential centrifugation
Protease-wash
Rate-zonal gradient centrifugation Isopycnic gradient centrifugation

Method 1
SDS-PAGE/Trypsin/HPLC/Ion Spray/ MS-MS

Virus solubilzation:
+/- alkylation
Reduction
+/- phenol or phenol-toluene extraction
SDS/crystal violet/Ficoll
100 ^{o}C

Protein separation and fragmentation:
One-dimensional SDS-PAGE
Sypro-Ruby staining
Gel slices
Imbibe with trypsin
Eluted tryptic fragments

Peptide separation:
Tryptic fragments injected onto C-18 reverse phase LC

Mass spectromety:
Electrospray ionization injection
Q-TOF Ultima
MS/MS acquisitions – 60 to 1900 daltons

Mass ion analyses:
Masslynx produce peak lists
MASCOT to NCBI (nr database)
Mass accuracy at 0.1 daltons
Protein identification: $p < 0.5$

Relative abundance:
emPAI

Method 2
PPS/Trypsin/HPLC/MS-MS

Virus solubilization:
PPS
100 ^{o}C
Reduction
Alkylation

Protein fragmentation:
Trypsin
Acid to hydrolyse PPS

Peptide separation:
Tryptic fragments injected onto C-18 reverse phase LC

Mass spectromety:
LTQ-Velos or LCQ Deca XP Plus
MS/MS acquisitions – exclusion list 100 daltons

Mass ion analyses:
Xcalibur produced peak lists
X!Tandem and SPIRE to custom DB
Mass Accuracy at 2.5 Da
Protein identification <= 1% FDR

Figure 1

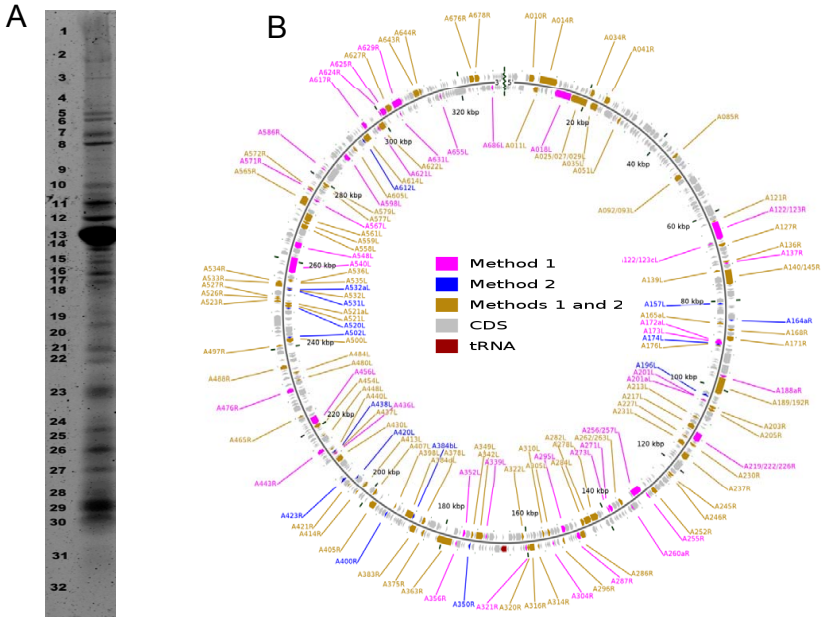
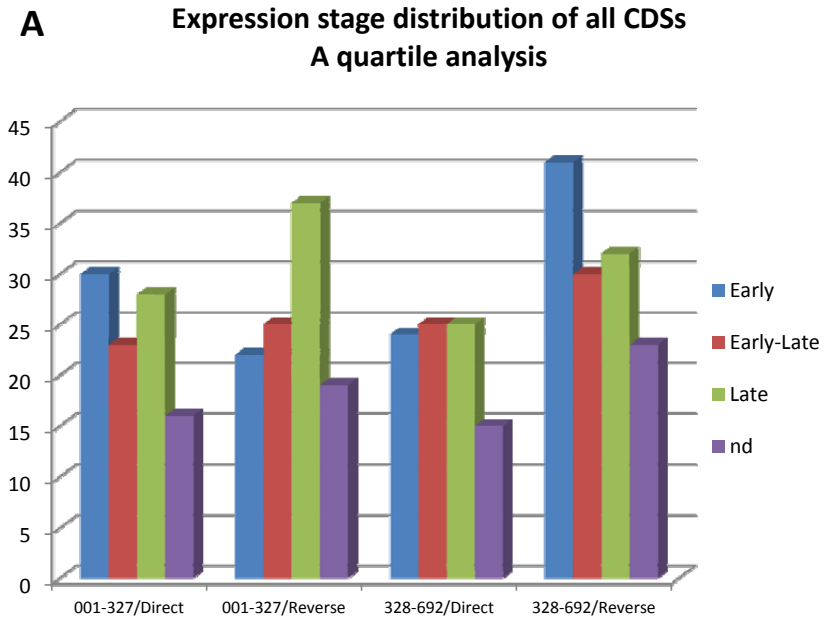


Figure 2.



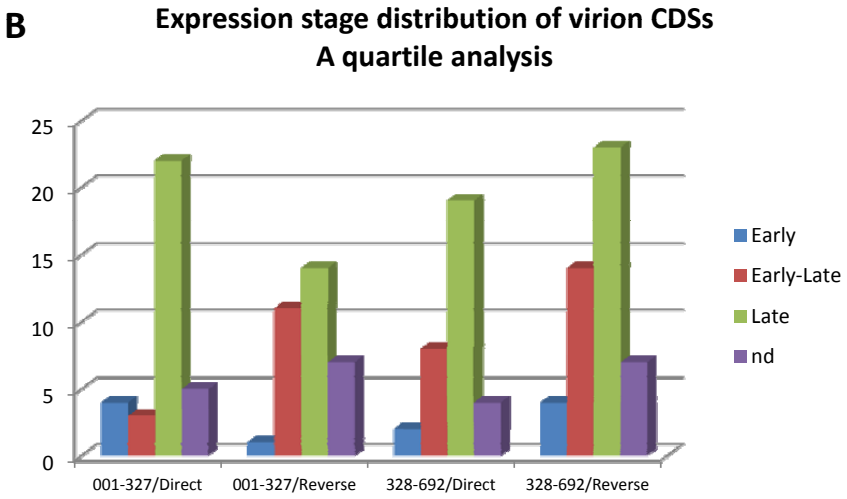


Figure 3

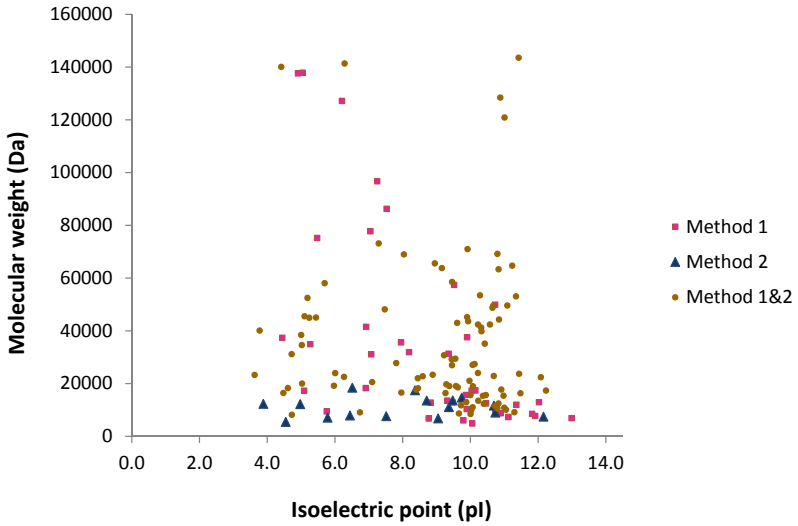
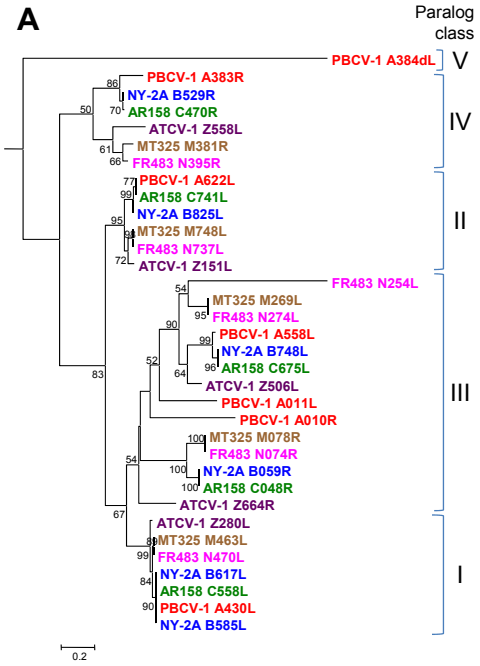


Figure 4



B

Domain and number of amino acids

Paralog class	PBCV-1 capsid proteins	A (D1)	B	C	D (D2)	E	Relative abundance (emPAI)	Estimated copies per virion#
I	A430L	195			241		13.1	1440
II	A622L	190	119		209		4.2	480
III	A010R	195			214		6.2	720
	A011L						13.7	1440
	A558L						6.0	720
IV	A383R	175		184	95		1.1	20
V	A384dL	135			119	45	1.0	20
		Chitin biinding Peritrophin-A domain			NCLDV superfamily capsid domain	Chitin biinding Peritrophin-A domain		

C

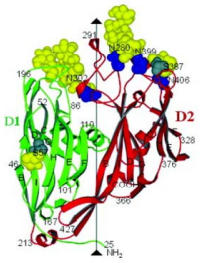


Fig. 5