The signature protein of the PVC superphylum

Ilias Lagkouvardos¹, Marc-André Jehl², Thomas Rattei³ and Matthias Horn#¹

¹Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria
²Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany
³Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

Running title: The PVC signature protein

# horn@microbial-ecology.net
Abstract

The phyla Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae, Poribacteria and OP3 comprise bacteria that share a common ancestor but show highly diverse biological and ecological features. Together they constitute the PVC superphylum. Using large scale comparative genome sequence analysis we identified a protein uniquely shared among all known members of the PVC superphylum. We provide evidence that this signature protein is expressed by representative members of the PVC superphylum. Its predicted structure, physicochemical characteristics, overexpression in *E. coli* and gel retardation assays with purified signature protein suggest a housekeeping function with unspecific DNA/RNA binding activity. Phylogenetic analysis demonstrated that the signature protein is a suitable phylogenetic marker for members of the PVC superphylum, and the screening of published metagenome data indicated the existence of additional PVC members. This study provides further evidence for a common evolutionary history of the PVC superphylum and presents a unique case in which a single protein serves as an evolutionary link between otherwise highly diverse members of major bacterial groups.
Introduction

The bacterial phyla *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and *Lentisphaerae* along with the candidate phyla *OP3* and *Poribacteria* were proposed to share a common ancestor based on their monophyletic grouping in 16S rRNA based phylogenetic trees (1). This diverse assemblage of phyla was termed PVC superphylum and later received additional support from genomic and phylogenetic analysis of conserved proteins (2-4). Most recently, 16 housekeeping and ribosomal proteins were used to infer evolutionary relationships among the PVC superphylum (5). This further established the common evolutionary origin of members of the PVC superphylum.

Despite their common origin, members of the PVC superphylum differ largely with respect to life style, physiology and ecology (1). Each phylum includes members that attracted significant research interest due to their importance in carbon and nitrogen cycling (e.g. *Rhodopirellula* and *Kuenenia* species; (6, 7)), as pathogens or symbionts (e.g. *Chlamydia* and *Protochlamydia* species; (8-10)), or as environmental microbes in aquatic and soil habitats (e.g. the *Verrucomicrobia*; (11, 12)). In addition to their ecological, biotechnological, and medical relevance, some members of the PVC superphylum show genetic and cellular features that are unusual for bacteria but reminiscent of eukaryotes or archaea (13-15). Due to these similarities members of the PVC superphylum have been implicated in the emergence and evolution of eukaryotes, a hypothesis that is controversially discussed (14, 16-20).

In this study we performed an extensive comparative genomic analysis in order to identify unifying links between the diverse members of the PVC superphylum. We describe the analysis and characterization of a protein, independently identified very recently (5), that is shared by all members of the superphylum but absent in all other bacteria. Computational analysis and functional assays provided
evidence for a putative house-keeping function of this protein. Due to its conservation among the PVC superphylum we were able to use this protein to extract information on the occurrence and diversity of members of the PVC superphylum in available environmental metagenomes.

**Materials and Methods**

*Identification of the signature protein*

Predicted coding sequences from completely sequenced PVC and representative non-PVC genomes were obtained from INSDC (21) and NCBI RefSeq (22) databases. All-vs.-all pairwise sequence similarities were pre-calculated by our SIMAP database (23). From SIMAP we obtained all bi-directionally best matching protein pairs (BBH) between all genomes, in which the alignment covered at least 50% of both protein sequences and the e-value was not higher than 1e-04. The score of each BBH was additionally used as threshold to determine inparalogs from the respective genomes. In order to cluster BBHs from the PVC superphylum into Clusters of Orthologous Groups (COGs) we first determined all 3-cliques (triangles) formed by PVC BBHs. Triangles were grouped into COGs if they shared a BBH. Remaining PVC BBHs were added to COGs if one of the proteins was already member of a COG and the other not. All other PVC BBHs were considered as individual COGs. Inparalogs associated to BBH proteins were added to the respective COGs in all clustering steps mentioned above.

For each COG we determined the presence/absence of the proteins encoded in PVC genomes. For COGs occurring in all PVC genomes, we determined the presence/absence in the representative non-PVC genomes from BBHs between PVC and non-PVC genomes. Only one COG, the PVC signature protein, was present in all PVC genomes and absent in all non-PVC genomes.

*COG based presence/absence analysis*
The COGs for all bacterial genomes were obtained from the eggNOG (24) database. The BBHs between PVC and non-PVC genomes described above were used to determine the presence/absence of each COG in the PVC genomes not yet contained in eggNOG. A matrix was then created with all COGs in the first column and the organisms name on top row. Then the table was filled with 1 or 0 for each COG for each genome based on presence or absence respectively allowing for quick overview of COG conservation across PVC and non PVC bacteria as selective sums.

For each COG without representatives in the PVC superphylum the *E. coli* representative was found and used as query in searches against Refseq database in NCBI (22) using BLAST (25). The ten first proteins of non-redundant origin (different organisms) were collected. Using these sets the average protein size and isoelectric point (pI) were calculated for each COG. The pI was calculated solving the Henderson-Hasselbach equation by a local Perl script.

**Screening of metagenome data**

All assembled metagenomes available at the JGI Genome Portal (26) were downloaded and organized into BLAST databases using makeblastdb (included in the BLAST+ suit) according to their originating environment. The nucleotide databases were searched for the presence of signature protein (SP) using tBLASTx (25) with default settings and all known signature protein sequences as queries. The output files were then merged and the matching translated sequences collected. All redundant (exact or substring match) and sequences containing stop codons or shorter than 45 amino acids were removed. The remaining sequences were submitted to the Conserved Domains Database CDD (27) and the presence of the SP domain was verified in all of them.
Phylogenetic analysis

Amino acid sequences from sequenced members of the PVC superphylum with or without metagenomic proteins were aligned using MUSCLE (28) in MEGA5 (29) and their evolutionary history was inferred using UPGMA (30) or FastTree (31). The evolutionary distances were computed using the JTT (32) for UPGMA and the WAG model (33) for FastTree, while a gamma value of 20 was used for both. Phylogenetic trees were visualized using iTOL (34).

Reverse transcriptase PCR

Verrucomicrobium spinosum DSM 4136 and Rhodopirellula baltica SH1 were inoculated from colonies grown on agar plates to flasks containing 100 ml of appropriate media described by Schlesner (35, 36), respectively, and grown while shaking at 22 °C. Initially, growth characteristics were determined measuring OD$_{600nm}$ in a spectrophotometer. Cultures were harvested after 3 d (exponential growth phase) and 5-6 d (stationary phase), respectively. Cells were lysed by bead beating (FastPrep FP120, SAVANT), and total RNA was extracted using Trizol (Molecular Research Center, Inc.) according to the manufacturer’s instructions. Primers were designed to target the genes encoding the V. spinosum and R. baltica SP, respectively (VssignF 5’-TCCCA GCATCGTAGTCTCAA-3’, VssignR 5’-TAAGCTTCCGGCTTGGTCT-3’and RbsignF 5’-TAAGAGTCGCAACGTCCTGA-3’, RbsignR 5’-TTCTTCTTGTGTCGGCTTC-3’). The housekeeping gene coding for glyceraldehyde 3-phosphate dehydrogenase (GAPDH) from V. spinosum was used as positive control (37) (VsqapdhF 5’-CGGTCTCTTTACCGAAGCTG -3’, VsqapdhR 5’-CGTTGGAGATGATGTTGTTG -3’). Reverse transcriptase PCR was performed using M-MLV polymerase (Invitrogen) and annealing temperatures of 55 °C for 35 cycles.
Cloning, expression, and purification of recombinant proteins

The genes coding for the SP of *R. baltica* (Genbank/EMBL/DDBJ accession number KF733603) and *P. amoebophila* (YP_008052) were synthesized (GenScript Corp. USA) flanked by restriction sites for *Eco*RI (Thermo Scientific) and *Xho*I (Thermo Scientific) that were used for subsequent cloning into pGEX 4T-1 vector (GE Healthcare) containing an N-terminal GST tag at the multiple cloning site. The final constructs were then transformed into electro-competent *E. coli* strain BL21 (λDE3). Transformed *E. coli* cells were grown overnight in 5 ml of Luria-Bertani (LB) medium containing 50µg/ml ampicillin (LB-Amp) at 37 °C on a shaker (120 rpm ) and the next day 1 ml from each culture were used to inoculate flasks containing 100 ml of LB-Amp. The cells were incubated for 2 hours (OD$_{600nm}$ ≈ 0.4) and then the expressions of the proteins were induced by 100 µM IPTG (final concentration). After 2 hours of induction cells expressing the GST-signature fusion protein were collected by centrifugation in 50 ml tubes at 6000 rpm for 10 min at 4°C. The supernatant was discarded and the tubes containing the cell pellets were stored at -20 °C.

For protein purification, the collected cell pellets were resuspended by vortexing in 4.5 ml binding buffer (125 mM Tris, 150 mM NaCl, 1% Triton X, and pH 8, with protease inhibitors (Roche Diagnostics) plus 0.5 ml lysozyme from a 2 mg/ml stock solution. The tubes were incubated for 15 min at room temperature on a rocking platform horizontally and then placed on ice. The final cell disruption was performed with three rounds of sonication for 30 sec at 70% strength (Bandelin Electronic) with intervals of cooling. The lysates were centrifuged at 12000 rpm for 10 min and the supernatant was transferred in new tubes. After three rounds of washing in 20 ml binding buffer, 2 ml glutathione coated magnetic beads slurry (Pierce), were mixed with the lysate and kept shaking horizontally for 1h. Using an appropriate magnetic stand, the beads were washed three times with washing buffer (125 mM Tris,
500 mM NaCl, 1% Triton X, and pH 8). Finally, 4 ml of elution buffer (125 mM Tris, 500 mM NaCl, 50 mM reduced glutathione, pH 9, with protease inhibitors was added and the beads were incubated for additional 15 min before elution, three times, keeping the eluates separated. The purity and quantity of purified proteins was determined by 12.5% SDS/PAGE and staining with colloidal Coomassie blue (Invitrogen).

For desalting, 2 ml of pooled protein purifications were placed in an Ultracell 10K spin column (Millipore) and PBS (pH 7.4) was used to fill the column up to 15 ml. The column was centrifuged for 30 min at 5000g. The desalting was repeated with another 15 ml of PBS resulting in 200µl of desalted and concentrated protein.

**Electrophoretic mobility shift assay**

For the evaluation of the effect of signature protein in the mobility of nucleic acids, purified proteins were mixed with DNA or RNA samples and gel loading dye (New England Biolabs). The mixtures were then loaded in 1% agarose gels and run for 1h under 120V and visualized by staining with ethidium bromide. When cleaved protein was used, 3µl of thrombin (GE Healthcare Life Sciences) were added to 30µl desalted and concentrated stock of fusion protein and left overnight at room temperature. Complete cleavage was verified then by SDS-PAGE.

**Results and Discussion**

To investigate the evolutionary history of the PVC superphylum, early after the original proposal we performed a comparative genome analysis to identify orthologous genes conserved among all PVC
members. We discovered a single protein coding gene of unknown function that is uniquely shared among all members of the superphylum that we refer to as the “signature protein” (SP) of the PVC superphylum (Lagkouvardos I., Rattei T. and Horn M., presented at the 8th German Chlamydia Workshop, Munich, 24 to 26 February 2010). In the following we verified its presence in all further sequenced PVC genomes published since using PSI-BLAST (25) and found the SP in all 55 available genome sequences. The only exceptions were: (i) missing gene predictions (e.g. for Rhodopirellula baltica SH1) that we identified only with tblastn, and (ii) incomplete genome sequences (e.g. the Poribacteria draft genome that has been estimated to represent 75% of the complete genome (38)) that we did not consider suitable for presence/absence analysis (Table S1). Recently, 16 housekeeping and ribosomal proteins were used to infer evolutionary relationships among the PVC superphylum (5), which further established the common evolutionary origin of the PVC superphylum. By searching for conserved signature insertions or deletions the same study independently recovered the SP to be encoded in all known members (except for the Poribacteria) (5). The presence of a protein in all PVC members that does not show any sequence similarity with other known proteins serves as a unifying link between the members of this diverse assemblage of microbes and suggests a conserved function.

Asking whether the SP is expressed, we searched available transcriptomic and proteomic data from members of the PVC superphylum. Members of the phylum Chlamydiae are best represented in such studies with few reports for Planctomycetes. We only found evidence for the expression in members of the Chlamydiae, where the SP seems to be expressed constitutively in low amounts similar to some house-keeping proteins (Table S2). To compensate for the lack of evidence of transcription for Planctomycetes and Verrucomicrobia, we performed reverse transcriptase PCR assays using RNA from Rhodopirellula baltica SH1 and Verrucomicrobium spinosum DSM 4136 isolated at logarithmic and
stationary growth phase, respectively. This demonstrated that the SP is expressed also in these organisms (Figure S1). Taken together, there is evidence for expression of the SP from representatives of all major phyla within the PVC superphylum.

The SP is a small, 50-60 aa protein exhibiting a considerable conservation in sequence (55% average amino acid sequence similarity among all representatives; Figure 1A) and physicochemical properties. In silico prediction of localization, isoelectric point and secondary structure revealed a highly basic cytosolic protein (pI 10-11) (39) consisting of an alpha helix followed by a putative second alpha helix depending on the prediction software (Figure 1B), which is reminiscent of the DNA binding helix-turn-helix motif (40). Structure prediction and physicochemical characteristics thus point towards a nucleic acid associated protein such as histone-like proteins, transcription factors, or ribosomal proteins. Consistent with this observation, the SP has been recognized as protein family in TIGRFAM (TIGR04137; (41)), where a possible ribosomal RNA interaction is proposed.

To verify the in silico prediction and to investigate the in vitro activity of the SP, we heterologously expressed the SPs of Protochlamydia amoebophila (as representative of the Chlamydiae) and Rhodopirellula baltica (Planctomycetes) as glutathione sulfotransferase (GST) fusion proteins in E. coli. The expressed proteins were purified with glutathione coated magnetic beads and subsequently used for gel retardation assays. When the fusion proteins were incubated with various DNA and RNA products (sheared genomic DNA, total RNA, or PCR products) this resulted in the retardation of nucleic acid mobility in agarose gels (Figures 1C). This was also observed after removal of the GST tag by protease treatment, but never when only the GST tag was used (Figures 1C, S2). A dose-dependent retardation was observed when increasing amounts of SP were added to PCR products (Figure 1C).
Together this demonstrates an unspecific and concentration dependent DNA and RNA binding activity of the SP from *R. baltica* and *P. amoebophila* *in vitro*. This mode of nucleic acid interaction seems to rule out a role of the SP as transcription factor, which typically show highly specific DNA binding activity.

To investigate whether the SP could function as a histone-like protein, we analysed *E. coli* cells overexpressing *R. baltica* or *P. amoebophila* SP. Overexpression of histones generally leads to nucleation of chromatin, which can be detected by staining with DNA-specific dyes (40). However, no nucleation was observed during over expression of both SPs in *E. coli* (Figure S3). Although we cannot exclude a histone function of the SP *in vivo* in *R. baltica* or *P. amoebophila*, expression in the heterologous host does not support such a role. In addition, *P. amoebophila* showing a condensed nucleoid in the elementary body stage encodes other histone-like proteins likely involved in chromatin condensation (42, 43).

The occurrence and documented expression of the SP in all members of the PVC superphylum point towards a highly conserved function. This function could be either unique to the superphylum, or the SP could substitute the role of an otherwise conserved and essential protein in non PVC organisms. To search for proteins that are well conserved in most other organisms but lack in PVC members we conducted a COG based comparative analysis among all available PVC genomes and a representative set of non-PVC genomes. This analysis revealed several highly conserved bacterial functions with no representation by a protein homolog in the PVC superphylum (Table 1). From those bacterial homologs missing in PVC bacteria, the ribosomal protein L30 that is also present in Archaea and Eukaryotes shows a striking physicochemical similarity to the SP of the PVC superphylum. Despite of the absence
of any amino acid sequence similarity, both proteins share similar size, pI, and expression profiles (Table 1). Together with its observed nucleic acid binding activity this suggests the possibility of the SP being a functional analog of the ribosomal protein L30 that is missing in all members of the superphylum. Further experimental investigation is needed to verify the presence and function of the SP in the ribosome of PVC members.

The high sequence conservation and exclusive presence of the SP in all members of the PVC superphylum suggests that it may serve as an additional phylogenetic marker for the superphylum. In fact, the topology of amino acid based phylogenetic trees resembles that of the 16S rRNA gene (Figure S4). Simple clustering using UPGMA recovered all PVC phyla with good bootstrap support, and the structure within the different phyla is largely similar (Figure S4A). The 16S rRNA tree topology was less well recovered in approximately maximum-likelihood SP trees using FastTree (31). Here the Verrucomicrobia SPs were not monophyletic, but included the Lentisphaerae sequences (Figure S4B). Still, the overall congruence between 16S rRNA gene and SP based trees allowed us to exploit the SP for the analysis of metagenomic datasets from various environmental samples to obtain insights into the diversity of the PVC superphylum. To this end, metagenomic datasets available in IMG/m (44) and SIMAP (45) were first screened using PSI-BLAST (46). In addition, tblastx was used to detect the SP even in the absence of correctly predicted coding sequences. In total 233 non-redundant SP sequences were detected, mainly in metagenomes originating from freshwater (36%), soil (34%), and marine (21%) samples (Table S1). Phylogenetic analysis of these sequences showed that the majority of metagenomic SPs are related to one of the known phyla within the PVC superphylum (Figure 2). Within the different phyla, however, several novel evolutionary lineages could be observed significantly expanding the known diversity of the PVC superphylum as inferred from SP phylogeny. The
Lentisphaerae was the least diverse phylum, followed by the Chlamydiae and OP3; the Planctomycetes and the Verrucomicrobia were most diverse. Interestingly, the majority of Verrucomicrobia sequences originated from soil metagenomes, while most OP3 and Chlamydiae sequences originated from freshwater samples (including sediments); no trend was observed for the other phyla. Although this analysis cannot be used to quantitatively assess abundance of PVC microbes in the different habitats, the observed ecological patterns are consistent with those of known members of the superphylum. For example, the relatively low number of metagenomic SPs branching with known members of the Chlamydiae phylum is consistent with the generally low abundance of chlamydial protein sequences detected in metagenomes in a recent study (47). An explanation for this could be the low abundance of Chlamydiae (which are typically associated with eukaryotic hosts) in environmental samples, which would result in a low coverage of chlamydial genomes in metagenomic datasets. Overall, the suitability of the SP as phylogenetic marker allows to identify genomic fragments containing the SP as originating from a PVC member, and thus helps in binning metagenomic data and in estimating the overall presence of PVC members in such datasets. In addition, concatenation of the SP with other conserved proteins should help to construct robust phylogenetic trees to analyze diversity and evolutionary history of the PVC superphylum (5).

In summary, all known members of the PVC superphylum share a small, conserved signature protein with nucleic acid binding activity. There is evidence for the expression of this protein for some PVC members, and its physicochemical properties, structure predictions, and the absence of ribosomal protein L30 in all members of the superphylum suggests that the SP has a conserved function and is possibly associated with the ribosome. We demonstrated that the SP is a useful marker for the analysis
of metagenomic data and may serve to investigate the diversity and ecology of bacteria related to this medically and biotechnologically important superphylum.

Acknowledgement

This work was funded by Austrian Science Fund (FWF) grant Y277-B03 and the University of Vienna (Graduate School “Symbiotic Interactions”). Matthias Horn acknowledges support from the European Research Council (ERC StG “EvoChlamy”).

References


Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2013. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. Isme J.


Table 1: Conserved functional categories among bacterial genomes absent in members of the PVC superphylum. COGs absent in all members of the PVC superphylum but conserved in at least 60% of all non-PVC bacteria analysed are listed together with basic physicochemical properties. The SP of the PVC superphylum is shown for comparison.

<table>
<thead>
<tr>
<th>COGs</th>
<th>PVC (56)</th>
<th>Other (490)</th>
<th>Function</th>
<th>Average length [aa]</th>
<th>Average pI</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG0806</td>
<td>0</td>
<td>466</td>
<td>16S rRNA processing protein RimM</td>
<td>182</td>
<td>4.8</td>
</tr>
<tr>
<td>COG0779</td>
<td>0</td>
<td>439</td>
<td>ribosome maturation factor RimP</td>
<td>154</td>
<td>4.5</td>
</tr>
<tr>
<td>COG1559</td>
<td>0</td>
<td>405</td>
<td>aminodeoxychorismate lyase</td>
<td>339</td>
<td>8.9</td>
</tr>
<tr>
<td>COG1841</td>
<td>0</td>
<td>375</td>
<td>ribosomal protein L30/L7E</td>
<td>60</td>
<td>11.0</td>
</tr>
<tr>
<td>COG1660</td>
<td>0</td>
<td>334</td>
<td>predicted P-loop-containing kinase</td>
<td>294</td>
<td>5.8</td>
</tr>
<tr>
<td>COG2884</td>
<td>0</td>
<td>327</td>
<td>cell division ATP-binding protein FtsE</td>
<td>224</td>
<td>9.5</td>
</tr>
<tr>
<td>COG2177</td>
<td>0</td>
<td>318</td>
<td>cell division protein FtsX</td>
<td>304</td>
<td>7.6</td>
</tr>
<tr>
<td>COG0595</td>
<td>0</td>
<td>300</td>
<td>mRNA degradation ribonuclease J1/J2</td>
<td>537</td>
<td>5.4</td>
</tr>
<tr>
<td>SP</td>
<td>56</td>
<td>0</td>
<td>DNA/RNA binding</td>
<td>60</td>
<td>11.6</td>
</tr>
</tbody>
</table>
**Figure legends**

**Figure 1:** Features of the PVC superphylum signature protein. (A) Conservation of the SP amino acid sequence. A sequence logo based on a MUSCLE alignment of all known SP generated by WebLogo 3 is shown (28, 48). The overall height of the alignment positions indicates the sequence conservation, while the height of symbols indicates the relative frequency of each amino acid at the respective position. Symbol colors reflect amino acid chemical properties. Highly conserved positions can be observed along the complete length of the alignment, with a longer conserved region in the middle, corresponding to a predicted α-helix.

(B) Predicted secondary and tertiary structures of representative signature proteins compared to small DNA/RNA binding proteins of *E. coli*. Predictions were performed using I-TASSER (49)(i-iii) and QUARK server (50) (iv-vi). (i, iv) SP of *Protochlamydia amoebophila UWE25* (Genbank/EMBL/DDBJ accession number YP_008052), (ii, v) SP of *Verrucomicrobium spinosum* (WP_009960041), (iii, vi) SP of *Rhodopirellula baltica* (KF733603), (vii) *E. coli* ribosomal protein L30 (PDB accession number 2AW4), (viii) *E. coli* DNA-binding protein H-NS (1HNS), (ix) *E. coli* histone like protein HU (1MUL).

Pink, alpha-helix; yellow, beta-sheet; blue, turn; grey, unstructured. Independent of the software a central alpha-helix is predicted for all SPs. The SP of all Planctomycetes shows a C-terminal, lysine rich extension which forms additional secondary structure elements.

(C) Nucleic acid mobility retardation by SP of *R. baltica* and *P. amoebophila*. (i) retardation assay with sheared genomic DNA. 1, molecular weight marker; 2, empty lane; 3, genomic DNA with GST tagged SP of *R. baltica*, 4: GST tagged SP of *R. baltica* without DNA, 5: genomic DNA only. (ii) retardation assay with purified total RNA. 1,5, molecular marker; 2, RNA only; 3, RNA with GST tagged SP of *R. baltica*.
baltica; 4, RNA with GST tagged SP of P. amoebophila; 6, RNA with GST only. Arrows indicate bands representing the 16S and 23S rRNA, respectively. (iii) retardation assay with PCR products. 1, molecular marker; 2, only PCR product; 3-7, PCR product with increasing concentration of GST tagged SP of P. amoebophila; 8, PCR product with GST only. The same molecular weight marker was used in all experiments, fragment sizes in bp are indicated. The retardation assays suggest an unspecific binding of SP to DNA and RNA.

Figure 2. Evolutionary relationships of all known PVC superphylum signature proteins and their metagenomic homologs. The environmental origin of SPs is color-coded at the tips of the tree for metagenomic sequences but not for SPs originating from complete genome sequences. An approximate maximum-likelihood tree is shown; nodes with less than 70% support are collapsed.