The ISME Journal (2011) 5, 1178–1190

www.nature.com/ismej

ORIGINAL ARTICLE

Comparative metagenomics of microbial traits within oceanic viral communities

Itai Sharon[1,2], Natalia Battchikova[3], Eva-Mari Aro[3], Carmela Giglione[4], Thierry Meinnel[4], Fabian Glaser[5], Ron Y Pinter[2], Mya Breitbart[6], Forest Rohwer[7,8] and Oded Béjà[1]

[1]*Faculty of Biology, Technion, Israel Institute of Technology, Haifa, Israel;* [2]*Faculty of Computer Science, Technion, Israel Institute of Technology, Haifa, Israel;* [3]*Department of Biochemistry and Food Chemistry, Molecular Plant Biology, University of Turku, Turku, Finland;* [4]*CNRS, Institut des Sciences du Végétal, Gif/Yvette cedex, France;* [5]*Bioinformatics Knowledge Unit, Lorry I Lokey Interdisciplinary Center for Life Sciences and Engineering, Technion, Israel Institute of Technology, Haifa, Israel;* [6]*College of Marine Science, University of South Florida, St Petersburg, FL, USA;* [7]*Department of Biology, San Diego State University, San Diego, CA, USA and* [8]*Center for Microbial Sciences, San Diego State University, San Diego, CA, USA*

Viral genomes often contain genes recently acquired from microbes. In some cases (for example, *psbA*) the proteins encoded by these genes have been shown to be important for viral replication. In this study, using a unique search strategy on the Global Ocean Survey (GOS) metagenomes in combination with marine virome and microbiome pyrosequencing-based datasets, we characterize previously undetected microbial metabolic capabilities concealed within the genomes of uncultured marine viral communities. A total of 34 microbial gene families were detected on 452 viral GOS scaffolds. The majority of auxiliary metabolic genes found on these scaffolds have never been reported in phages. Host genes detected in viruses were mainly divided between genes encoding for different energy metabolism pathways, such as electron transport and newly identified photosystem genes, or translation and post-translation mechanism related. Our findings suggest previously undetected ways, in which marine phages adapt to their hosts and improve their fitness, including translation and post-translation level control over the host rather than the already known transcription level control.

The ISME Journal (2011) **5,** 1178–1190; doi:10.1038/ismej.2011.2; published online 10 February 2011
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: cyanophage; gene transfer; metagenomics; photosynthesis; viral–host interactions

Introduction

Microbial genes acquired by viruses presumably improve virus fitness (Lindell *et al.*, 2005; Dammeyer *et al.*, 2008) and may also contribute to the increasing host range (Rohwer and Thurber, 2009). The role of these genes may be significant both from the evolutionary point of view in processes such as horizontal gene transfer, as well as in their contribution to global processes such as photosynthesis (Sharon *et al.*, 2007). Recent studies have demonstrated the acquisition of microbial genes by different marine phages (viruses that infect bacteria), including genes for phosphate metabolism (Rohwer *et al.*, 2000; Sullivan *et al.*, 2005; Martiny *et al.*, 2009; Kathuria and Martiny, 2011), pigment biosynthesis and photosynthesis (Mann *et al.*, 2003; Lindell *et al.*, 2004; Millard *et al.*, 2004; Sullivan *et al.*, 2005; Dammeyer *et al.*, 2008; Sharon *et al.*, 2009; Alperovitch *et al.*, 2010), as well as different electron transport proteins (Lindell *et al.*, 2004; Sullivan *et al.*, 2005; Alperovitch *et al.*, 2010). Recently, gene cassettes containing whole photosystem-I (PSI) gene suites sufficient to build a monomeric PSI were also reported to exist in marine cyanophages (viruses that infect cyanobacteria; Sharon *et al.*, 2009; Alperovitch *et al.*, 2010). These cassettes were detected using a specific PSI-targeted search scheme using both the Global Ocean Survey (GOS) dataset (Rusch *et al.*, 2007) and viral and microbial marine biomes from the Pacific Northern Line Islands (Dinsdale *et al.*, 2008a, b). Despite their importance, these instances may be regarded as anecdotal and provide only a partial view of the phenomenon. The extent to which viral acquisition of microbial genes is prevalent is yet to be determined.

In this work, we have studied evidence for viral acquisition of microbial genes in marine environments. A semiautomatic computational pipeline was designed in order to achieve this goal, which takes advantage of publically available databases

Correspondence: O Béjà, Faculty of Biology, Technion, Israel Institute of Technology, Haifa 32000, Israel.
E-mail: beja@tx.technion.ac.il
Received 23 August 2010; revised 4 November 2010; accepted 13 December 2010; published online 10 February 2011

such as RefSeq and non-redundant of the National Center for Biotechnology Information, as well as 454-pyrosequencing generated databases from marine environments. We have used the pipeline to study microbial genes that were acquired by viruses in the GOS scaffold database and that were not previously detected in viral genomes in general.

Materials and methods

The general scheme used for finding viral scaffolds containing microbial genes previously undetected on fully sequenced viral genomes is presented in Figure 1. In this study, we describe in details the different steps.

Collecting potential viral GOS scaffolds
As a first step, a set of GOS scaffolds containing at least one viral-like gene was collected (Figure 1, box 1). The process was carried out using the reciprocal best blast hit approach: first, all 75 079 proteins available in the National Center for Biotechnology Information RefSeq-viral database (April 2009; Pruitt *et al*., 2007, 2009) were blasted (−p tblastn −F F −e 1e−20) against the GOS scaffolds dataset. Overall, a total of 223 687 scaffolds containing viral-like regions were collected in this stage. Next, all regions matching one of the RefSeq-viral proteins in these scaffolds were blasted (−p blastx −F F −e 1e−10) against a unified database of RefSeq-viral and RefSeq-microbial proteins. Scaffolds with microbial best hit were excluded from further analysis leaving 84 921 scaffolds with at least one viral-like gene.

Scaffold annotation and gene contents-based filtering
All scaffolds containing at least one viral-like gene were annotated using an iterative blast procedure (Figure 1, box 2). The process begins by blasting the scaffolds against a combined database of RefSeq-viral and RefSeq-microbial (−p blastx −e 1e−3 –F F). For each scaffold, information for the best alignment is recorded, including the best RefSeq hit's id, position on the scaffold, percent identity and *e*-value. Next the corresponding region on the scaffold is masked by a stretch of N's and the process continues with the modified scaffolds. A scaffold that did not get any hit is removed from further annotation and the process is repeated with the remaining scaffolds until no more scaffolds are left. Overall, 202 176 regions similar to 15 900 RefSeq proteins from the unified database were annotated. Of all RefSeq hits, 4169 originated from RefSeq-viral and 11 731 from RefSeq-microbial. We refer to this set of RefSeq proteins as the set of proxy proteins.

On the basis of gene annotation, all scaffolds that fulfilled the following criteria were considered to be viral.

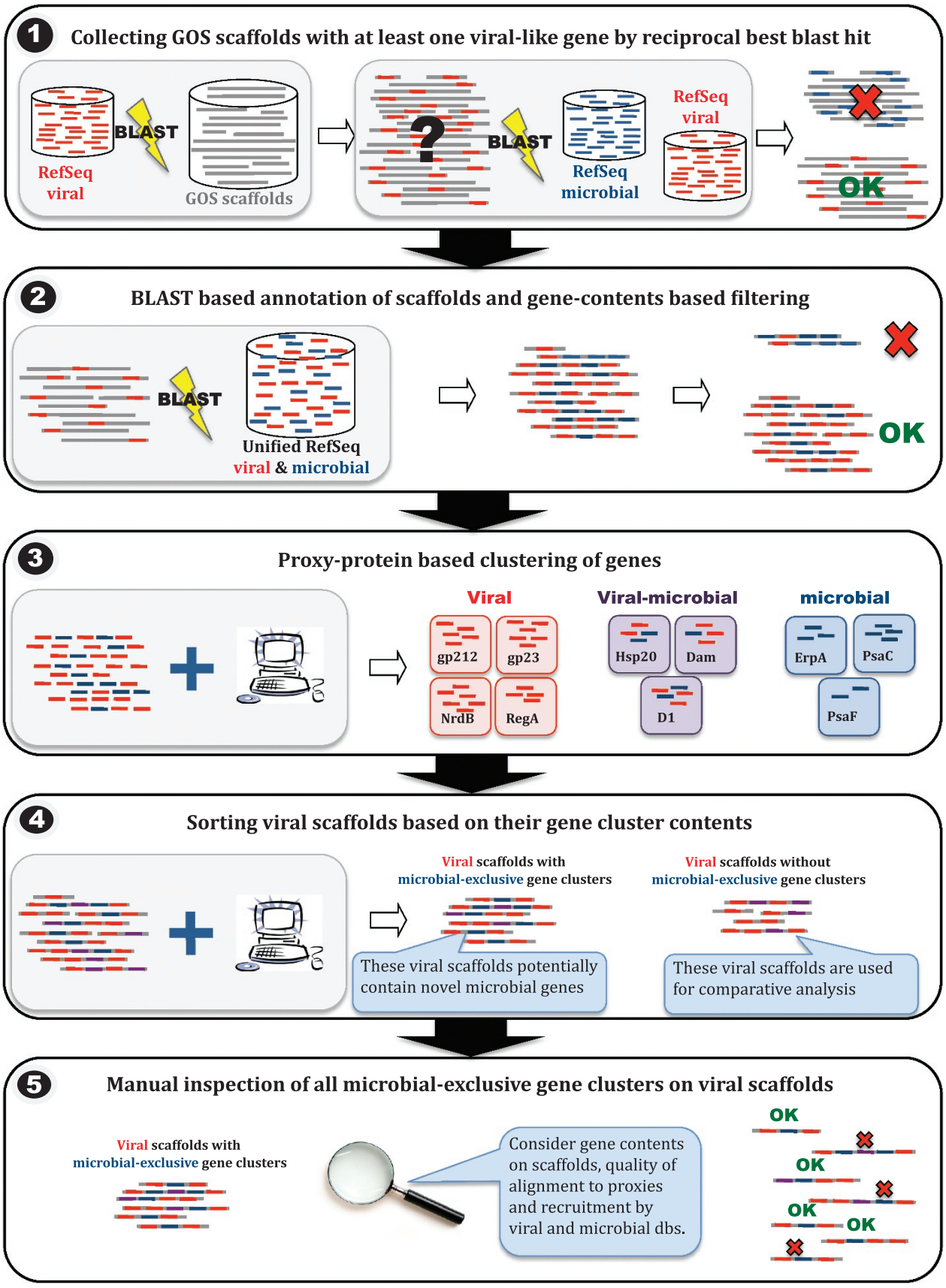
- Scaffolds containing three genes, in which the proxy proteins for the two edge genes are of viral origin.
- Scaffolds containing four genes or more, in which the number of genes with viral proxy proteins is at least three and their share is at least 20%.

Scaffolds that did not fulfill the above criteria were excluded from further analysis.

Gene clustering
Metagenomic data is fragmented by its nature, and therefore even similar genes extracted from metagenomic data may not share the same region and as a result may not be recognized as homologs (Figure 1, box 3). To overcome this obstacle genes in our dataset were clustered based on sequence similarity between their proxy proteins. We used a simple graph-based clustering algorithm that identifies components in the graph, in which each node (proxy protein) is connected by its edges (an edge connects nodes representing proxy proteins with high sequence similarity) to at least 70% of all other nodes in the component (refer to the Supplementary Information for a complete description). This strict criterion was meant to minimize wrong clustering of proxy proteins on account of possible redundancy in clusters representing similar functions. Such redundancies were partially resolved by semantic clustering, as well as manual inspection of resulting clusters. In all, 12 257 (89%) of the 13 741 clusters that were found consisted of a single proxy protein. Largest cluster, in terms of number of its proxy proteins, is composed of 37 proxy proteins, all of them are PsbA proteins from different sources (refer to Supplementary Table S1 for a list of clusters with the largest numbers of proxy proteins).

A cluster may be composed of proxy proteins that are either viral exclusive (namely were seen in the context of viral genomes only, for example, NrdA and NrdB), microbial exclusive (for example, PsaA and PsaB) or viral-microbial (for example, D1 and D2). Classification of each proxy protein was determined by blasting it against the other RefSeq database than the one it was taken from (−p blastp −e 1e−3 −F F). Overall, 1840 proxy proteins had homologs in RefSeq-viral alone, 5829 had homologs in RefSeq-microbial alone and 8231 had homologs in both databases. Next, each cluster was tagged as either viral exclusive, microbial exclusive or viral-microbial based on the following criteria.

- Viral exclusive—all members are from RefSeq-viral and have homologs only in RefSeq-viral.
- Microbial exclusive—all members are from RefSeq-microbial and have homologs only in RefSeq-microbial.
- Viral-microbial—some members have homologs in both RefSeq-viral and RefSeq-microbial.


1
Collecting GOS scaffolds with at least one viral-like gene by reciprocal best blast hit
BLAST
RefSeq viral
GOS scaffolds
?
BLAST
RefSeq viral
RefSeq microbial
OK
2
BLAST based annotation of scaffolds and gene-contents based filtering
BLAST
Unified RefSeq viral & microbial
OK
3
Proxy-protein based clustering of genes
Viral
gp212
gp23
NrdB
RegA
Viral-microbial
Hsp20
Dam
D1
microbial
ErpA
PsaC
PsaF
4
Sorting viral scaffolds based on their gene cluster contents
Viral scaffolds with microbial-exclusive gene clusters
These viral scaffolds potentially contain novel microbial genes
Viral scaffolds without microbial-exclusive gene clusters
These viral scaffolds are used for comparative analysis
5
Manual inspection of all microbial-exclusive gene clusters on viral scaffolds
Viral scaffolds with microbial-exclusive gene clusters
Consider gene contents on scaffolds, quality of alignment to proxies and recruitment by viral and microbial dbs.
OK
OK
OK
OK

Clusters having both viral- and microbial-exclusive members and no viral-microbial members are considered to be conflicting clusters. In this case, the core set is split into two sub-core sets, one viral and one microbial. Only about 1% of the clusters were identified as conflicting.

Cluster annotation was determined based on the majority of its member annotations. All clusters having identical annotation and the same tagging were unified (semantic clustering) to generate the final set of clusters. This set was used in the next two steps for searching for new microbial-exclusive functions (defined in terms of gene clusters) on viral genomes, whereas viral-exclusive and viral-microbial clusters were used for validation and context analysis.

Collecting viral scaffolds with microbial clusters

All scaffolds containing at least one member of a microbial cluster were considered to be of interest (Figure 1, box 4). Overall, 3046 scaffolds containing at least one gene that belongs to one of 200 microbial clusters with four members or more were detected. A second set of scaffolds representing viral scaffolds with no microbial-exclusive clusters was created for comparative analysis purposes. All scaffolds carrying at least two genes with at least 60% (majority) of their genes having proxy proteins from RefSeq-viral that did not contain any microbial exclusive clusters were considered in this set. These criteria are conservative and were chosen, as no manual inspection of these scaffolds was carried out. Overall, 39 866 scaffolds containing 111 825 genes of viral or viral-microbial clusters were found.

Manual inspection of microbial clusters on viral scaffolds

Each of the 200 microbial clusters on viral genomes has gone through manual inspection in order to determine both the credibility of its viral origin as well as its annotation (Figure 1, box 5). Criteria in this case are more loose, as certain parameters may compensate for other. Specifically, the following parameters were considered.

Cluster annotation parameters

- Percent identity of cluster members to their proxy proteins. Clusters with percent identity $>50\%$ between cluster members and their proxy proteins were considered to be good, 30–50% is marginal, whereas $<30\%$ is poor. Inclusion/exclusion of marginal clusters was carried out based on this and other annotation-related measures (see below), whereas clusters with poor percent identity were excluded from further analysis.
- Consistency of start/end coordinates in the alignments of cluster members and their proxy proteins. For example, we would expect a gene located in the middle of a scaffold to cover its proxy protein throughout most of its length. Gene parts located at the scaffold edges or near a gap should match either the beginning or end parts of their proxy proteins depending on their location.
- Possible alternative annotations to the cluster members. This was carried out by searching for more similar proteins in the non-redundant database of the National Center for Biotechnology Information.
- Possible wrong alignment due to low-complexity regions.

Viral affiliation of scaffolds containing microbial clusters

- Recruitment against Northern Line Islands viral and microbial biomes (Dinsdale *et al.*, 2008a, b). Recruitment is defined as the fraction of a scaffold that was aligned against at least one read with percent identity of at least 85% on at least 80% of the read's length. High recruitment from the viral samples combined with low recruitment from the microbial samples supports the affiliation of a scaffold as viral, whereas the opposite suggests a possible wrong affiliation. Zero or close to zero recruitment from both biomes does not support or object any affiliation.
- Fraction of genes on scaffolds whose proxy proteins are viral. We expect to find at least some of each cluster members on scaffolds enriched with genes with viral proxy proteins.

Figure 1 Pipeline for the identification of viral scaffolds with microbial genes. (**1**) Candidate scaffolds containing at least one potential viral gene were identified using a bi-directional BLAST (Basic Local Alignment Search Tool) search of all proteins in the RefSeq-viral database of the National Center for Biotechnology Information against all GOS scaffolds, and then a BLAST search of all significant best hits against a combined RefSeq-viral and RefSeq-microbial proteins database. Scaffolds with best hits from RefSeq-viral were further considered. (**2**) Gene contents of all candidate scaffolds were determined using an iterative sequence similarity-based procedure against the combined dataset of RefSeq-viral and RefSeq-microbial proteins. (**3**) Genes were clustered based on sequence similarity of their RefSeq hits, with each cluster being tagged based on the origin of its RefSeq proteins (viral exclusive, microbial exclusive or viral microbial). For the purpose of clustering we consider all genes' proxy proteins rather the genes and their scaffolds. Scaffold tagging was determined based on gene contents and on recruitment against the Northern Line Islands datasets (Figure 2). (**4**) Viral scaffolds were sorted into scaffolds containing members from viral or viral-microbial clusters only, and scaffolds also containing members of microbial exclusive clusters. Purple genes refer to genes that belong to viral-microbial clusters, red and blue refer to members of viral-exclusive and microbial-exclusive clusters, respectively. (**5**) The latter set was inspected manually in order to validate their origin and annotation.

- Consistency between the origins of viral proxy proteins of genes on the scaffold. Viral affiliation of scaffolds whose proxy proteins are from the same virus or viral group (for example, cyanophages) is considered to be more reliable.
- Percent identity of genes from viral clusters to their proxy proteins (similar criteria to the one listed earlier of microbial exclusive clusters).
- Possible alternative non-viral affiliation of the viral genes. Once again, this was carried out by searching for other proteins from non-redundant that are similar to the viral genes on the scaffolds.

Overall, 34 microbial gene clusters containing at least four members were detected on 452 scaffolds. In this paper, we do not discuss clusters with hypothetical genes; only gene clusters with annotations.

Phylogenetic tree reconstruction
Phylogenetic trees were constructed using the FastTree program (Price *et al.*, 2010) with multiple alignments produced by MUSCLE (Edgar, 2004). Although using MUSCLE default parameters, FastTree was run using parameters for high accuracy: -spr 4 (to increase the number of rounds of minimum-evolution SPR moves) and -mlacc 2 -slownni (to make the maximum-likelihood nearest neighbor interchanges search more exhaustive). These parameters can produce slight increases in accuracy. To estimate the reliability of each split in the tree, FastTree uses the Shimodaira–Hasegawa test on the three alternate topologies (nearest neighbor interchanges) around that split (Guindon *et al.*, 2009).

Results and discussion

Identification of viral scaffolds containing microbial genes
To study microbial genes acquired by viruses, we devised a general and broad pipeline for identifying scaffolds of viral origin that contain microbial genes in the GOS metagenomic dataset (Rusch *et al.*, 2007). The GOS dataset consists of long reads and scaffolds, making it ideal for this kind of analysis. The large amount of viral genes found in the GOS dataset in previous studies (Williamson *et al.*, 2008; Comeau *et al.*, 2010) suggests that a significant portion of the GOS sequences are of viral origin. However, these studies focused on gene families rather than whole scaffolds, and were thus limited to known viral genes only. The pipeline developed for this study consists of five stages (Figure 1, refer to the Materials and methods section for a detailed description), and was designed specifically to identify viral scaffolds containing genes previously detected on microbial genomes only. Sequence similarity-based identification and annotation of viral scaffolds (boxes 1 and 2 in Figure 1) were performed using the National Center for Biotechnology Information's RefSeq-viral (Pruitt *et al.*, 2009) and RefSeq-microbial protein databases. These databases consist of known proteins from all fully sequenced viral and microbial genomes, and can thus provide the annotation of genes and tag them as either viral or microbial. Scaffold tagging (box 2, Figure 1) was determined based on gene content and recruitment against the Northern Line Islands viral and microbial datasets (Dinsdale *et al.*, 2008a, b; Sharon *et al.*, 2009). As demonstrated in Figure 2a, the recruitment analysis provides strong evidence that the scaffolds, which were selected based on their gene content, are indeed viral (refer also to Supplementary Figure S3 in the Supplementary information). All genes on viral scaffolds were clustered based on their best hits from the unified RefSeq-viral and RefSeq-microbial dataset (Figure 1, box 3). Clusters were tagged as viral exclusive, microbial exclusive or viral-microbial, based on their members (see Materials and methods section). Gene clusters containing only microbial genes (microbial exclusive) were considered to be of interest (Figure 1, box 4). These clusters were inspected manually in order to verify their credibility (Figure 1, box 5). Overall, a total of 34 clusters spanning 452 scaffolds of average length 2546 (minimum 879 bp, maximum 20 741 bp) passed through all the filters (Table 1).

Microbial gene families found on GOS viral scaffolds
Various microbial gene families were found via our search regime, including mainly energy and carbohydrate metabolism genes (see Kyoto Encyclopedia of Genes and Genomes; Kanehisa *et al.*, 2008; category enrichments; Figure 2b) and different case studies were chosen for further analysis and are presented here (the full list of resulting microbial genes and families may be found in Table 1 and searched for on the accompanying website http://www.cs.technion.ac.il/~itaish/VirMic/).

Antioxidation genes. One of the most abundant viral gene clusters (35 incidences) was related to antioxidant proteins from the peroxiredoxin protein family (also referred to as AhpC/Tsa in bacteria), a ubiquitous family of proteins that use sulfhydryl groups to remove peroxides (clusters T768 and T712). Most of the peroxiredoxin genes detected appear on scaffolds with a suspected cyanophage origin. In all, 33 out of 35 cases are from cyanophages based on neighboring genes on the scaffold; in fact, using our selection scheme, 92% of the viruses found to contain microbial genes were most likely cyanophages (Supplementary Figure S5). This deviates slightly from the 76% of cyanophages present in the set of all of viral scaffolds in GOS (see Supplementary Information and Supplementary Figure S4). Interestingly, the phage's

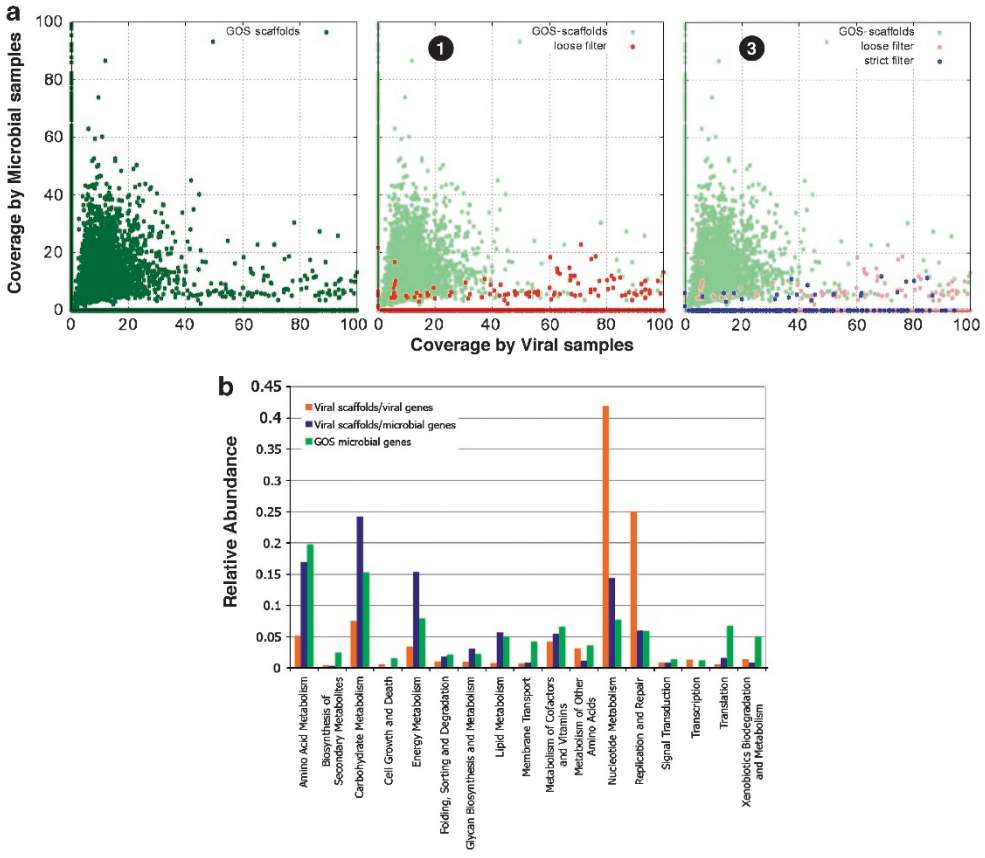


Figure 2 Enrichment scheme for microbial traits in viral landscapes. (**a**) Recruitment coverage of GOS scaffolds enriched through the different selection filters with Northern Line Islands marine viromes (*x*-axis) and microbiomes (*y*-axis; Dinsdale *et al*., 2008a, b). Left panel, previous selection; middle panel, loose filter settings corresponding to box 1 in Figure 1; right panel, strict filter settings corresponding to box 2 in Figure 1. Coverage is defined as percentage of GOS scaffold length covered by at least one recruited read. (**b**) Distribution of viral (red) and microbial (blue) genes on viral scaffolds, as well as genes selected at random from other GOS scaffolds (green). Kyoto Encyclopedia of Genes and Genomes categories 'energy metabolism' and 'carbohydrate metabolism' are enriched in the microbial genes/viral scaffolds set with respect to the GOS gene set (*P*-values of 5e−10 and 4e−9, respectively). Refer to the Supplementary Information for a full description.

peroxiredoxin proteins had higher identity to peroxiredoxin proteins from alphaproteobacteria (~60% on the protein level) than to peroxiredoxin proteins from cyanobacteria (<50% on different marine *Synechococcus* or *Prochlorococcus* peroxiredoxin proteins). A protein-based phylogenetic tree (Supplementary Information, Figure S1A) also reveals that most viral peroxiredoxins form distinct clusters separate from the cyanobacterial peroxiredoxins. The viral peroxiredoxins contained both conserved cysteine residues (see Supplementary Information, Figure S1B for protein alignment) important for the catalytic cycle of the peroxide reaction in 2-Cys peroxiredoxins (Hall *et al*., 2009). Peroxiredoxins were recently suggested to have a role in oxygenic photosynthesis, as cyanobacteria contain larger peroxiredoxin gene families than non-photosynthetic bacteria (Tripathi *et al*., 2009). It is, therefore, tempting to suggest that the phages accumulate these peroxiredoxin genes from different bacterial sources in order to enhance the defense against potential oxidative damage during oxygenic photosynthesis. Alternatively, these genes could have been originated from cyanobacterial hosts but modified extensively by the phages to such extent that the signal for their origin was lost.

Translation and co-translation mechanism genes. Phages would benefit greatly from being able to control host gene expression. It is, therefore, interesting that the phages in our clusters carry modified versions of predicted proteins homologous to proteins involved directly in translation, such as ribosomal protein S21 (cluster T72), translation initiation factor IF-1 (cluster T255) or proteins involved in co- and post-translational modification, such as phosphorylases (cluster T891) or peptide deformylases (cluster T255). Interestingly, peptide deformylases (PDFs) represent the most abundant family found by our search scheme with 70 cases. Only one case of viral PDF has been identified to date (Seguritan *et al*., 2003). PDFs are enzymes responsible for the irreversible co-translational removal of the formyl moiety from the *N*-formyl methionine found at the beginning of all prokaryotic

Table 1 Manually validated microbial gene clusters on viral scaffolds with at least four instances

Cluster	*Size*	*Annotation*
T128	70	Peptidedeformylase
T1396	58	Glycerol-3-phosphate cytidylyltransferase
T982	36	Exported protein
T451	31	Glycinedehydrogenase
T768	30	Antioxidant, AhpC/Tsa family protein
T1414	28	Fructose-1,6-bisphosphate aldolase class I
T100	20	Glycosyltransferase, group 1
T156	15	Serine hydroxymethyltransferase
T212	13	NAD(P)H-dehydrogenase subunit D
T603	13	NAD(P)H dehydrogenase subunit I
T17	13	Membrane protein
T90	13	Phosphoribosylaminoimidazole-succinocarboxamidesynthase
T1338	12	Glycosyltransferase involved in cell wall biogenesis-like
T1486	11	NH(3)-dependent NAD+ synthetase NadE
T71	10	Glycine cleavage system protein P2
T885	10	Iron–sulfur cluster-binding protein
T891	8	Phosphorylase
T327	8	Mannitol-1-phosphate/altronate dehydrogenase
T198	8	Alkylhydroperoxidereductase/thiol-specific antioxidant/Mal allergen
T1113	8	Putative NH(3)-dependent NAD synthetase
T596	8	Scaffold protein
T158	7	Sugarisomerase
T255	6	Translation initiation factor IF-1
T1119	6	Coenzyme PQQ biosynthesis protein A
T445	6	Adenylatekinase
T444	5	Iron–sulfur cluster insertion protein ErpA
T883	5	Glycine cleavage system aminomethyltransferase T
T712	5	Antioxidant, AhpC/TSA family protein
T446	5	Photosystem II reaction center protein N
T361	5	5′-Methylthioadenosine nucleosidase/*S*-adenosylhomocysteinenucleosidase
T72	5	Ribosomal protein S21
T204	4	Photosystem I reaction center subunit IX (PsaJ)
T1557	4	Photosystem I subunit VII (PsaC)
T109	4	Photosystem I P700 chlorophyll a apoprotein A1 (PsaA)

Refer to the accompanying website (http://www.cs.technion.ac.il/~itaish/VirMic) for a full description of each cluster.

translation products (that is, bacteria and organelles). More than 98% of prokaryotic proteins undergo removal of their formyl group, a reaction that is needed to unmask the first methionine. The first methionine is in turn also most often co-translationally removed by another enzyme called methionine aminopeptidase. The removal of the formyl group and the first methionine is an essential and ubiquitous process called N-terminal methionine excision; indeed, both PDF and methionine aminopeptidase genes are part of the minimal genome required for the viability of bacteria (Giglione *et al.*, 2004). PDF genes could be identified in almost all genomes, including both eukaryotes and prokaryotes (Giglione *et al.*, 2004), and are divided into three types (type I, type II and type III, see Supplementary Information, Figure S2) based on protein homology and structural similarities. Subtype I*B* refers to both plastidial and bacterial PDFs. The C-terminal helix of subtype I*B* PDFs displays no conservation in length (Giglione *et al.*, 2009) and is not required for catalytic activity (Meinnel *et al.*, 1996). Recently, it has been proposed that bacterial subtype I*B* PDFs might interact with the ribosome via their C-terminal helix extension (Bingel-Erlenmeyer *et al.*, 2008; Giglione *et al.*, 2009; Kramer *et al.*, 2009). In contrast, for variants with a shortened C-terminal helix, it has been suggested that other domains may be required for correct ribosome binding (Giglione *et al.*, 2009). Most interestingly, although few of the viral PDFs contain a modified C-terminal helix extension, we reveal that most viral versions of the PDFs do not contain a C-terminal extension at all (Figure 3). The C-terminally truncated viral PDFs belonging to subtype I*B* clearly constitute a new branch in the PDF tree (Supplementary Information, Figure S2), including two sub-branches, one more closely related to cyanobacteria and the other to plastidial PDF from the marine unicellular microeukaryotes (Figure 3a). Complete absence of the C-terminus, as described above, has been reported in only a few PDFs that often belong to a family of inactive type III PDFs. Nevertheless, it should be stressed that the absence of deformylase activity in some type III PDFs is because of only a single substitution at the active site that is unmodified in viral PDFs, that is, the crucial Gln of motif 1, and not to the absence of the C-terminus (Figure 3). From a global comparison with other known PDFs, it is highly likely that phage

Figure 3 Marine viral PDFs do not display the conserved C-terminal helix of subtype IB PDF and constitute a new class of active PDFs. (**a**) A phylogenetic tree was constructed from an alignment of ~200 PDFs recapitulating sequence and phylogenetic diversity (Supplementary Figure S2). Among marine viral PDFs, only representative members are displayed (colored in gray). These proteins are related to PDFs from cyanobacteria (shown in green) and photosynthetic planktonic picoeukaryotes (shown in blue). Proteins showing the closest similarities around the three conserved motifs 1, 2 and 3 (colored in yellow) and C-helix were selected and realigned, and the motifs are shown. Unlike type III inactive PDFs (colored in orange below), all required residues of the motifs are conserved, which is strongly suggestive of peptide deformylase activity. The closest structural models (that is, *E. coli* and *Arabidopsis thaliana* PDF1B) are indicated on top. In both cases, the C-terminus folds as a α helix. (**b**) A refined three-dimensional model for viral PDF (top) compared to thethree-dimensional crystal structure of the most relevant PDF (see panel **a**) from chloroplastic PDF1B (PDB code 3cpmA; bottom). The two structures are shown in the same orientation, that is, toward the entry of the active site crevice. Both N- and C-ends are indicated in white with N and C, respectively.

PDFs correspond to the first naturally occurring active PDFs devoid of the C-terminal domain. A three-dimensional model strongly favors this hypothesis (Figure 3b). These newly identified viral PDFs are predicted to bind directly to the elongating nascent chains or to the productive ribosome via another, yet unidentified dedicated ribosome-binding domain, without the need for a C-terminal α-helix. This could involve new interactions rendered possible by the genuine ribosomal proteins brought by the viral genome.

In bacteria, the PDF-Met:tRNA^{fMet} formyltransferase operon, which encodes the two genes involved in both formylation of the initiator tRNA and deformylation of the nascent peptides, is not subject to metabolic control, a mechanism coupling the concentration of components of the biosynthetic mechanism to growth rate (Meinnel and Blanquet, 1993). Instead, saturation of the N-terminal methionine excision system in bacteria by protein overproduction, such as that originating from viral particle assembly, results in the production of incorrectly processed or unprocessed proteins that might be poorly active or less accumulated. This issue is well illustrated in the case of chloroplast photosystem II proteins, such as D1 (PsbA) and D2 (PsbD), whose accumulation strongly depends on deformylation status (Giglione *et al.*, 2003).

Evidence that viral proteins do undergo both deformylation and N-terminal methionine removal with rules similar to that of the bacterial host is available from several viral types (Beyreuther and Gronenborn, 1976; Shank *et al.*, 1977; Sauer and Anderegg, 1978; Aono *et al.*, 1982; Maley *et al.*, 1983). The potential of newly identified PDFs may suggest a paradigm, in which phages control both viral and host protein expression in a competitive and/or synergic manner, working on translational and particularly co-translational levels, as opposed to the known phage-based control at the transcription level (Courey, 2001; Paul, 2008).

Energy metabolism genes

Carbon metabolism genes. We identified 28 cases of fructose bisphosphate aldolase (FBA) class I genes (all on scaffolds with a suspected cyanophage origin), which form distinct clades from the cyanobacterial FBA-I clade. FBA is a core carbon metabolic enzyme responsible for the aldol cleavage of fructose-bisphosphate into glyceraldehyde-3-phosphate and dihydroxyacetone phosphate sugars in glycolytic reactions, and reverse aldol condensation of these triose sugars to fructosebisphosphate in Calvin cycle and gluconeogenic reactions. FBA exists in two non-homologous but functionally equivalent forms, referred to as class I FBAs (FBA-I) and class II FBAs (FBA-II; Marsh and Lebherz, 1992). Although cyanobacteria usually encode for FBA-II, marine *Synechococcus* and *Prochlorococcus* encode for both FBA-I and FBA-II (some *Prochlorococcus* strains, NATL1A and NA-

TL2A, encode for FBA-I only; Rogers *et al.*, 2007). The observation of FBA-I in cyanophages is of special interest as it has been reported that several cultured phages carry genes for TalC. The phage TalC version could potentially function as a fructose-6-phosphate aldolase or as a transaldolase. Currently, based on protein alignment, the cyanophage TalC is suspected of acting as a transaldolase and being involved in metabolizing carbon substrates during host infection (Sullivan *et al.*, 2005).

NAD(P)H dehydrogenase genes. We found two high-confidence viral clusters, T212 and T603, related to type I NAD(P)H dehydrogenase or complex I. This enzyme transfers electrons from NADH via FMN and Fe–S centers to a ubiquinone molecule with concomitant pumping of protons across the inner membrane, thus catalyzing the first step in the mitochondrial electron transport chain. In cyanobacteria, type I NAD(P)H dehydrogenases or the NDH-1 complexes, participate in respiration, cyclic electron flow around PSI and CO_2 uptake. However, the nature of the electron donor (NADH, NADPH or ferredoxin), as well as the complete electron pathway in cyanobacterial NDH-1 remain uncertain (Battchikova and Aro, 2007).

Type I NAD(P)H dehydrogenases are multisubunit enzymes. Their composition varies between 13 and 15 subunits found in the bacterial NDH-1 complexes, and 45 subunits discovered in bovine mitochondrial complex I (Battchikova and Aro, 2007) and subunit nomenclatures differing among organisms. The T603 and T212 clusters are related to NdhI/NuoI/Nqo9/TYKY (cyanobacteria/*Escherichia coli*//*Thermus thermophilus*/*Bos taurus*) and NdhD/NuoM/Nqo13/ND4, respectively. Both subunits have been found in all known type I NAD(P)H dehydrogenases, including mitochondrial complex I. Conservation of these two subunits during evolution implies their importance for the biological function of the enzyme.

As the absolute majority of microbial clusters on viral scaffolds originate from cyanophages, hereafter the cyanobacterial nomenclature (NdhI and NdhD) will be used for the sake of clarity. The T603 cluster proteins showed a significant homology to the cyanobacterial NdhI subunit, which is one of the Fe–S proteins in the hydrophilic domain of the NDH-1 complex and is directly involved in electron transfer. The crystal structure of the hydrophilic domain of the NDH-1 complex from *T. thermophilus* (Sazanov and Hinchliffe, 2006) demonstrated that the corresponding subunit Nqo9 contains two 4Fe–4S clusters, N6a and N6b, which constitute part of the main electron transfer route within the enzyme. The cysteine residues coordinating N6a and N6b in the cubane geometry are conserved in viral proteins of the T603 cluster (Figure 4a). Similar viral *ndhI* genes were recently detected in marine viral enriched fractions; within phages containing both PSI and photosystem-II genes (Alperovitch *et al.*, 2010).

The T212 cluster proteins are related to NdhD, a large hydrophobic subunit located in the membrane domain of the NDH-1 complexes. In some cyanobacterial species, small families of NdhD proteins have been discovered. NdhD1 genes have been found in all cyanobacteria sequenced to date, whereas the presence of other NdhD genes is more variable (Ogawa and Mi, 2007). Reverse genetics (Klughammer *et al.*, 1999; Ohkawa *et al.*, 2000) and proteomic analyses (Zhang *et al.*, 2004) demonstrated that NDH-1 complexes containing NdhD1 or NdhD2 proteins participate in respiration and cyclic electron flow around PSI, whereas NDH-1 complexes with NdhD3 and NdhD4 are responsible for low CO_2-inducible or constitutive CO_2 uptake, respectively (for reviews, see Battchikova and Aro, 2007; Ogawa and Mi, 2007). Aligning T212 cluster proteins with NdhD1–NdhD4 of *Thermosynechococcus* 7002 and *Synechocystis* 6803 showed a higher homology of viral proteins to NdhD1 or NdhD2 compared with NdhD3 or NdhD4 (Figure 4b). This implies a preference of cyclic electron flow and respiration during viral infection over inorganic carbon acquisition, thus probably enhancing the production of adenosine triphosphate for cyanophage reproduction.

Genes related to the assembly of Fe–S clusters. Fe–S proteins have a crucial role in electron transfer. We observed several viral protein clusters that are part of Fe–S complexes or are related to the assembly of Fe–S complexes. Three viral clusters were found relevant to the assembly pathways of Fe–S centers in bacterial proteins. Two occurrences of SufE protein genes were observed (T399). SufE together with SufS forms a complex that functions as a cysteine desulfurase (Ayala-Castro *et al.*, 2008). This enzyme works at the first step of the Fe–S cluster formation, liberating sulfur atoms from free cysteine. At the second step, liberated sulfur is donated to a protein, which functions as a scaffold for nascent Fe–S cluster assembly. Two clusters were found to correspond to scaffold proteins: T596 (U-type) and T444 (A-type). Following assembly on the scaffold, the Fe–S cluster is then transferred to a target apoprotein (Ayala-Castro *et al.*, 2008). This adds to the currently known viral proteins involved in electron transfer processes, the cyanophage-encoded plastocyanins and ferredoxins (Lindell *et al.*, 2004; Sullivan *et al.*, 2005).

Photosystem genes. Of the various photosynthesis genes identified by our search, two were new genes never before reported in viral genomes: (i) viral versions of the photosynthesis *psbN* gene (cluster T-446), which was only recently annotated correctly in the genome of myophage P-SSM4 (Sullivan *et al.*, 2010); and (ii) a PSI single copy of the *psaJ* gene (cluster T-204). The latter observation is of special

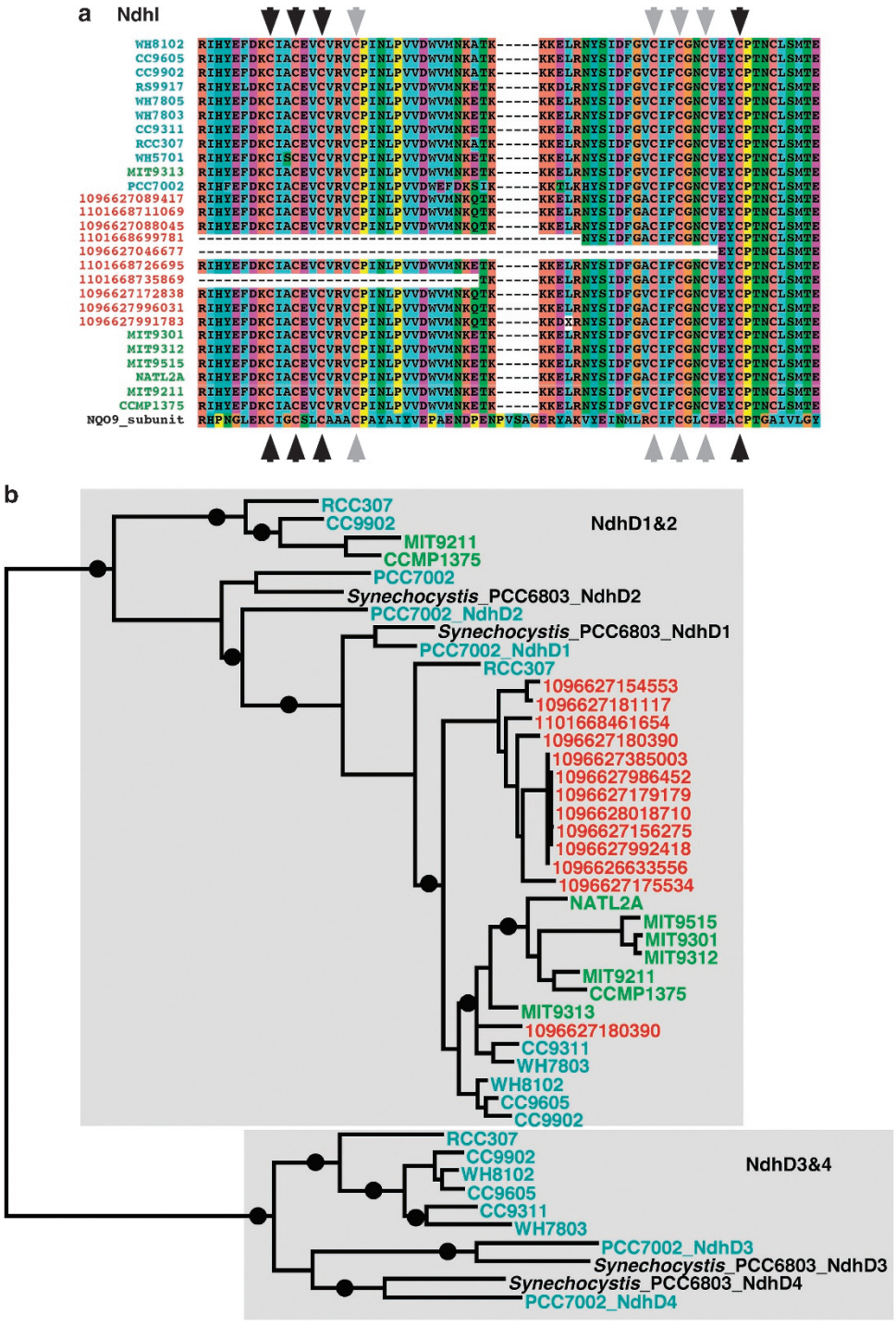


Figure 4 Viral NAD(P)H dehydrogenase subunits. (**a**) NdhI protein alignment. *Synechococcus* sequences are colored in cyan, *Prochlorococcus* in green and viral proteins in red. The Nqo9 sequence from *T. thermophilus* is shown for reference. The conserved cysteine residues coordinating Fe–S clusters in Nqo9 are marked with arrows (black for N6a and gray for N6b). For clarity, only part of the protein length is shown. (**b**) NdhD FastTree approximated maximum-likelihood phylogenetic tree. *Synechococcus* NdhD sequences are colored in cyan, *Prochlorococcus* in green and viral proteins in red. *Synechocystis* PCC6803 and *Synechococcus* PCC7002 NdhD1, 2, 3 and 4 sequences were used as references. Only bootstrap values above 80% are shown as black circles on the branches.

interest, as previously identified versions of viral *psaJ* genes were always found fused with *psaF* genes, together forming unique viral *psaJF* fusion genes (Sharon *et al*., 2009). Aligning the new viral PsaJ peptides with cyanobacterial peptides or with the PsaJ portion of PsaJF viral proteins (Figure 5) shows that it most closely resembles PsaJ peptides from *Prochlorococcus*. And yet, the viral PsaJ peptides have certain amino acids that are unique and are not found in *Prochlorococcus* peptides. The viral *psaJF* gene is the first in the viral PSI cassette order (*psaJF*→*C*→*A*→*B*→*K*→*E*→*D*). It was

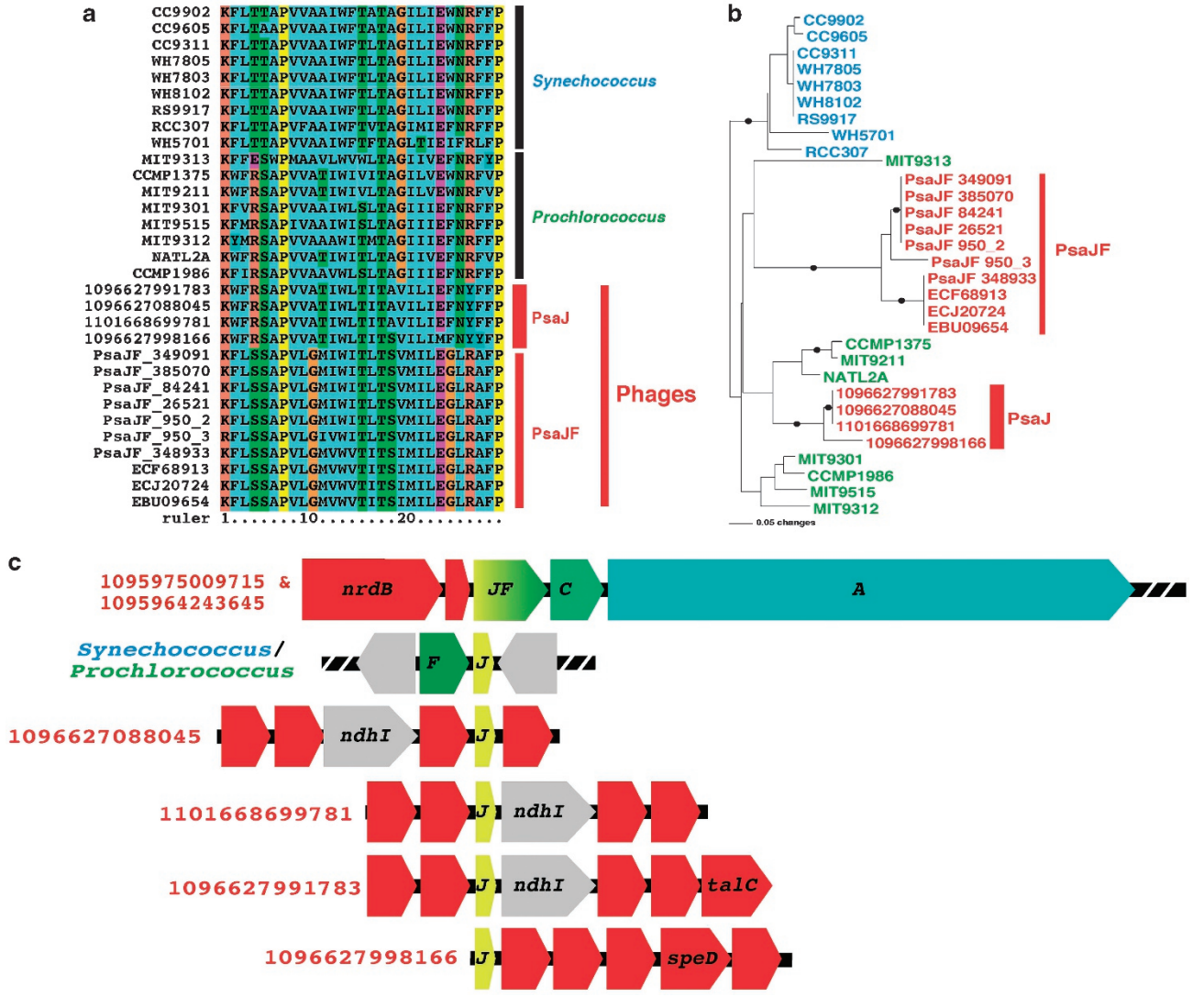


Figure 5 Viral PSI PsaJ protein. (**a**) PSI PsaJ protein and peptide alignment. Viral PsaJ proteins identified in this study and viral PsaJF (only the N-terminus, which contains the PsaJ portion, is shown) are labeled in red to the right of the alignment. (**b**) PsaJ FastTree approximated maximum-likelihood phylogenetic tree. *Synechococcus* PsaJ sequences are colored in cyan, *Prochlorococcus* in green, viral PsaJF and the newly identified viral PsaJ in red. Only bootstrap values above 80% are shown as black circles on the branches. (**c**) Schematic physical maps of selected viral clones or *Prochlorococcus* and *Synechococcus* genome fragments containing the PSI *psaJ* gene and a viral GOS clone containing the *psaJF* fusion gene. *ndhI* denotes NAD(P)H dehydrogenase I gene, *speD* a polyamine biosynthesis gene, *talC* a transaldolase gene and *nrdB* the ribonucleoside-diphoshate reductase-β 2 gene. Red-arrowed boxes mark viral genes and gray mark bacterial genes. PSI genes are colored in yellow, green and blue.

hypothesized that this unique cassette order is the result of many trial-and-error events performed by the cyanophages. It is, therefore, tempting to suggest that the event observed of a single *psaJ* in viral genomes is a snapshot of the first event in the evolution process of the construction of a full viral PSI cassette.

Conclusions

Significance of methodology

Our approach differs from previous studies of viral genes in the GOS dataset (Williamson *et al.*, 2008; Comeau *et al.*, 2010) by focusing on whole scaffolds rather than on protein families. Gene affiliation in this study was determined based on the affiliation of its scaffold, rather than its similarity to already known viral genes as was previously. This approach made it possible to reveal new gene families that were not previously known. Interestingly, >75% of the viral scaffolds that contained microbial-exclusive genes identified in the current study also contained viral genes that were identified by a previous study (Williamson *et al.*, 2008). Owing to the differences between the methodologies used in the two studies, we believe that this fact increases the credibility of our findings.

Implications

The mosaic theory of phage evolution (Hendrix *et al*., 1999) predicts that a global pool of genes exists that may recombine among different viruses to produce novel genotypes. Frequently genes that appear to originate from very divergent sources are incorporated into phage genomes. These genes are referred to as MORONs and may arise from both the global virome and microbial hosts. The latter possibility has profound implications on horizontal gene transfer and global distributions of genes. Genomic sequences of phages and giant viruses such as mimivirus show that viral genomes can accumulate MORONs, encoding a range of functions from modified photosynthetic reaction centers to aminoacyl tRNA synthetases. Comparisons of co-occurring viromes and microbiomes suggest that viruses are major pools of genomic diversity of core and specialized metabolisms (Dinsdale *et al*., 2008b). In fact, previous cumulative data suggest that the viruses represent the largest pool of genomic diversity on the planet (Rohwer, 2003; Angly *et al*., 2006). Furthermore, our results indicate that the oceanic virome is an almost unlimited source of naturally bioengineered genes.

The results presented here expand our knowledge of this viral-encoded gene pool. Identified viral auxiliary-enriched genes included genes for energy and carbohydrate metabolism, a significant deviation from the metabolic profile of known viral genes in GOS (Williamson *et al*., 2008), which consists for the most part of replication and repair genes and nucleotide metabolism genes. Nevertheless, the profile of acquired genes also deviates from the general profile of microbial genes in GOS (Yooseph *et al*., 2007), suggesting that only genes that are beneficial for viruses are acquired. Furthermore, our findings of potential phage control over the host via the translation level (as observed in some animal viruses) calls for concentrated efforts to isolate these yet uncultured phages.

Acknowledgements

We thank D Lindell for her encouragement and ideas. We are indebted to B Gronenborn (CNRS, Gif, France) for stimulating discussions. This work was supported in part by grant 1203/06 from the Israel Science Foundation (OB), the CoE project 118637 from the Academy of Finland (E-MA), CNRS grant PEPS-2008 and ARC (Association pour la Recherche contre le Cancer, Villejuif) grant 4920 (CG and TM), and grant DBI-0850206 from the National Science Foundation (MB).

References

Alperovitch A, Sharon I, Rohwer F, Aro E-M, Glaser F, Milo R *et al.* (2010). Reconstructing a puzzle: existence of cyanophages containing both photosystem-I & photosystem-II gene suites inferred from oceanic matagenomic datasets. *Environ Microbiol* **13**: 24–32.

Angly F, Felts B, Breitbart M, Salamon P, Edwards R, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Aono J, Kangawa K, Matsuo H, Horiuchi T. (1982). Coliphage 434 tof protein: NH2-terminal amino acid sequence and kinetic and equilibrium measurements of DNA binding. *Mol Gen Genet* **186**: 460–466.

Ayala-Castro C, Saini A, Outten FW. (2008). Fe-S cluster assembly pathways in bacteria. *Microbiol Mol Biol Rev* **72**: 110–125, table of contents.

Battchikova N, Aro E-M. (2007). Cyanobacterial NDH-1 complexes: multiplicity in function and subunit composition. *Physiol Plant* **131**: 22–32.

Beyreuther K, Gronenborn B. (1976). N-terminal sequence of phage lambda repressor. *Gen Genet* **147**: 115–117.

Bingel-Erlenmeyer R, Kohler R, Kramer G, Sandikci A, Antolic S, Maier T *et al.* (2008). A peptide deformylase-ribosome complex reveals mechanism of nascent chain processing. *Nature* **452**: 108–111.

Comeau AM, Arbiol C, Krisch HM. (2010). Gene network visualization and quantitative synteny analysis of more than 300 marine T4-like phage scaffolds from the GOS metagenome. *Mol Biol Evol* **27**: 1935–1944.

Courey AJ. (2001). Cooperativity in transcriptional control. *Curr Biol* **11**: R250–R252.

Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenberg-Dinkel N. (2008). Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18**: 442–448.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008a). Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L *et al.* (2008b). Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* **3**: e1584.

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

Giglione C, Boularot A, Meinnel T. (2004). Protein N-terminal methionine excision. *Cell Mol Life Sci* **61**: 1455–1474.

Giglione C, Fieulaine S, Meinnel T. (2009). Cotranslational processing mechanisms: towards a dynamic 3D model. *Trends Biochem Sci* **34**: 417–426.

Giglione C, Vallon O, Meinnel T. (2003). Control of protein life-span by N-terminal methionine excision. *EMBO J* **22**: 13–23.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* **537**: 113–137.

Hall A, Karplus PA, Poole LB. (2009). Typical 2-Cys peroxiredoxins: structures, mechanisms and functions. *FEBS J* **276**: 2469–2477.

Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* **96**: 2192–2197.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M *et al.* (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**: D480–D484.

Kathuria S, Martiny AC. (2011). Prevalence of a calcium-based alkaline phosphatase associated with the marine cyanobacterium *Prochlorococcus* and other ocean bacteria. *Environ Microbiol* **13**: 74–83.

Klughammer B, Sültemeyer D, Badger MR, Price GD. (1999). The involvement of NAD(P)H dehydrogenase subunits, NdhD3 and NdhF3, in high-affinity CO_2 uptake in *Synechococcus* sp. PCC7002 gives evidence for multiple NDH-1 complexes with specific roles in cyanobacteria. *Mol Microbiol* **32**: 1305–1315.

Kramer G, Boehringer D, Ban N, Bukau B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nat Struct Mol Biol* **16**: 589–597.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–89.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101**: 11013–11018.

Maley GF, Guarino DU, Maley F. (1983). Complete amino acid sequence of an allosteric enzyme, T2 bacteriophage deoxycytidylate deaminase. *J Biol Chem* **258**: 8290–8297.

Mann NH, Cook A, Millard A, Bailey S, Clokie M. (2003). Bacterial photosynthesis genes in a virus. *Nature* **424**: 741.

Marsh JJ, Lebherz HG. (1992). Fructose-bisphosphate aldolases: an evolutionary history. *Trends Biochem Sci* **17**: 110–113.

Martiny AC, Huang Y, Li W. (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340–1347.

Meinnel T, Blanquet S. (1993). Evidence that peptide deformylase and methionyl-tRNA(fMet) formyltransferase are encoded within the same operon in *Escherichia coli*. *J Bacteriol* **175**: 7737–7740.

Meinnel T, Lazennec C, Dardel F, Schmitter JM, Blanquet S. (1996). The C-terminal domain of peptide deformylase is disordered and dispensable for activity. *FEBS Lett* **385**: 91–95.

Millard A, Clokie MRJ, Shub DA, Mann NH. (2004). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101**: 11007–11012.

Ogawa T, Mi H. (2007). Cyanobacterial NADPH dehydrogenase complexes. *Photosynth Res* **93**: 69–77.

Ohkawa H, Pakrasi HB, Ogawa T. (2000). Two types of functionally distinct NAD(P)H dehydrogenases in *Synechocystis* sp. strain PCC6803. *J Biol Chem* **275**: 31630–31634.

Paul JH. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J* **2**: 579–589.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. (2009). NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**: D32–D36.

Pruitt KD, Tatusova T, Maglott DR. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61–D65.

Rogers MB, Patron NJ, Keeling PJ. (2007). Horizontal transfer of a eukaryotic plastid-targeted protein gene to cyanobacteria. *BMC Biol* **5**: 26.

Rohwer F. (2003). Global phage diversity. *Cell* **113**: 141.

Rohwer F, Thurber RV. (2009). Viruses manipulate the marine environment. *Nature* **459**: 207–212.

Rohwer F, Segall AM, Steward G, Seguritan V, Breitbart M, Wolven F *et al.* (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. *Limnol Oceanogr* **45**: 408–418.

Rusch DB, Halpern AL, Heidelberg KB, Sutton G, Williamson SJ, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: I, the northwest Atlantic through the eastern tropical Pacific. *PLoS Biol* **5**: e77.

Sauer RT, Anderegg R. (1978). Primary structure of the lambda repressor. *Biochemistry* **17**: 1092–1100.

Sazanov LA, Hinchliffe P. (2006). Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*. *Science* **311**: 1430–1436.

Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM. (2003). Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. *J Bacteriol* **185**: 6434–6447.

Shank PR, Hutchinson III CA, Edgell MH. (1977). Isolation and characterization of the four major proteins in the virion of bacteriophage phiX174. *Biochemistry* **16**: 4545–4549.

Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N *et al.* (2009). Photosystem-I gene cassettes are present in marine virus genomes. *Nature* **461**: 258–262.

Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB *et al.* (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J* **1**: 492–501.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**: e144.

Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR *et al.* (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**: 3035–3056.

Tripathi BN, Bhatt I, Dietz KJ. (2009). Peroxiredoxins: a less studied component of hydrogen peroxide detoxification in photosynthetic organisms. *Protoplasma* **235**: 3–15.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al.* (2008). The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**: e1456.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Zhang P, Battchikova N, Jansen T, Appel J, Ogawa T, Aro E-M. (2004). Expression and functional roles of the two distinct NDH-1 complexes and the carbon acquisition complex NdhD3/NdhF3/CupA/Sll1735 in *Synechocystis* sp PCC 6803. *Plant Cell* **16**: 3326–3340.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)