The genome of a prasinoviruses-related freshwater virus reveals unusual diversity of phycodnaviruses

Hao Chen¹, Weijia Zhang¹,⁵, Xiefei Li¹, Yingjie Pan¹,³, Shuling Yan¹,⁴ and Yongjie Wang¹,²,³*¹

Abstract

Background: Phycodnaviruses are widespread algae-infecting large dsDNA viruses and presently contain six genera: Chlorovirus, Prasinovirus, Prymnesiovirus, Phaeovirus, Coccolithovirus and Raphidovirus. The members in Prasinovirus are identified as marine viruses due to their marine algal hosts, while prasinovirus freshwater relatives remain rarely reported.

Results: Here we present the complete genomic sequence of a novel phycodnavirus, Dishui Lake Phycodnavirus 1 (DSLPV1), which was assembled from Dishui Lake metagenomic datasets. DSLPV1 harbors a linear genome of 181,035 bp in length (G + C content: 52.7%), with 227 predicted genes and 2 tRNA encoding regions. Both comparative genomic and phylogenetic analyses indicate that the freshwater algal virus DSLPV1 is closely related to the members in Prasinovirus, a group of marine algae infecting viruses. In addition, a complete eukaryotic histone H3 variant was identified in the genome of DSLPV1, which is firstly detected in phycodnaviruses and contributes to understand the interaction between algal virus and its eukaryotic hosts.

Conclusion: It is in a freshwater ecosystem that a novel Prasinovirus-related viral complete genomic sequence is discovered, which sheds new light on the evolution and diversity of the algae infecting Phycodnaviridae.

Keywords: Phycodnaviridae, Prasinovirus, DSLPV1, Histone H3, Diversity

Background

The phycodnaviruses comprise genetically diverse, morphologically similar, large icosahedral (100 ~ 200 nm), and double strand DNA (180 ~ 560 kbp) viruses that infect eukaryotic algae from both fresh and marine waters, and presently contain six genera: Chlorovirus, Prasinovirus, Prymnesiovirus, Phaeovirus, Coccolithovirus and Raphidovirus [1]. The family of Phycodnaviridae is placed within a major, monophyletic assemblage of large eukaryotic DNA viruses, termed the nucleo-cytoplasmic large DNA viruses (NCLDV), which was recently proposed to group into a new viral order Megavirales [2, 3]. Accumulating evidence suggests that these algae-infecting large viruses are active players in the aquatic ecosystem [4–6].

The hosts of phycodnaviruses are abundant and widespread in the natural environments, while almost all phycodnaviruses, including members of the Coccolithovirus, Phaeovirus, Prasinovirus, Prymnesiovirus and Raphidovirus genera, infect marine algae except for chloroviruses (Chlorovirus) that target freshwater algae [7, 8]. The members of Prasinovirus are the most studied among these marine algae-infecting viruses because the hosts of prasinoviruses are worldwide distributed and play a central role in the oceanic carbon cycle [9]. All known prasinoviruses infect marine photosynthetic picoeukaryotic algae in the class Mamiellophyceae, which mainly contains three genera: Ostreococcus, Micromonas and Bathycoccus [10, 11]. Thus far, several prasinoviruses have been isolated and sequenced. For instance, the Ostreococcus virus OtV5 (Ostreococcus tauri virus 5), possessing 186,234 bp long linear genome, was isolated from the green alga O. tauri [12], and the Micromonas virus MpV1 (Micromonas sp.
RCC1109 virus 1) was isolated from *M. sp.* RCC1109 in eutrophic northwestern Mediterranean coastal lagoons with a genome size of 184,095 bp [13]. Interestingly, in our previous work of culture-independent metagenomic analyses, genomes of prasinoviruses-related viruses, Yellowstone Lake Phycodnaviruses (YSLPVs), were discovered in Yellowstone Lake [14], a freshwater lake in Yellowstone National Park, Wyoming.

In this study, we assemble the complete genome of a *Prasinovirus*-related novel large phycodnavirus from freshwater metagenomic datasets in Shanghai, China. Comparative genomic and phylogenetic analyses reveal unique features of this new member of *Phycodnaviridae* and shed light on the evolution, diversity and distribution of phycodnaviruses.

**Methods**

**Metagenomic data source**

The Dishui Lake (DSL) metagenomic datasets were generated with Illumina Miseq sequencer and subjected to a series of processes for quality controlling as described previously [15]. In brief, water samples were taken from neritic area of Dishui Lake (121°55′ 27.00″ N 30°53′ 56.00″ E) in 2013. Microbial biomasses were collected onto 0.22-μm membrane filters, from which over 6 μg genomic DNA was extracted and used for metagenomic sequencing. Raw reads were firstly analyzed with FastQC and was extracted and used for metagenomic sequencing. Raw reads were firstly analyzed with FastQC and NGS QC Toolkit prior to de novo sequence assembly. Detail information of the datasets was shown in Additional file 1: Table S1.

**Metagenomic sequence assembly of viral genome**

All high-throughput sequenced reads in the DSL metagenomic paired-end libraries were assembled into contigs by using Newbler v2.6 (Roche) with default parameters. Among 1949 contigs with the length of more than 10 kb, one was initially confirmed to be related to algal large viruses with BLASTP searching of NCBI nr database and subsequently used as the template in reference assembly as previously described [16]. Briefly, this contig was used as the reference sequence to which all reads in the DSL metagenomic paired-end datasets were assembled with a minimum overlap length of 25 bp and minimum overlap identity of 95%. The reference assembly was repeated until the assembled sequence stopped extending. All assemblies were performed with the bioinformatics software Geneious R9 (Biomatters, https://www.geneious.com).

**Genome analysis**

Geneious R9 software (Biomatters) was used to predict open reading frames (ORFs) by defining a start codon of ATG and a minimum 150 bp. Translated amino acid sequences were used to search (E-value < 10⁻³) for homologs in NCBI nr database by using the BLASTP program. One top hit to virus and/or non-virus was recorded. Functional annotation of ORFs was performed using the InterProScan program (http://www.ebi.ac.uk/Tools/pfa/interproscan/) [17], and conserved domain was determined with both the NCBI server and the HHpred server (http://toolkit.tuebingen.mpg.de/hhpred) [18]. Transfer RNA (tRNA) sequences were identified by using the tRNAscan-SE tool [19]. Repetitive sequences were detected with both the Geneious Pro software and the Reputer program (http://bibiserv.techfak.uni-bielefeld.de/reputer/) [20]. Whole genome alignment at the nucleotide level was performed with the MAUVE software [21].

**Phylogenetic analysis**

Maximum likelihood phylogenetic trees were reconstructed based on homologs of DNA polymerase B family (PolB), nicotinamide adenine dinucleotide (NAD)-dependent epimerase/dehydratase gene and histone H3 gene, respectively. Homolog counterparts from a range of representative viruses, bacteria and eukaryotes were downloaded from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein). Amino acid sequences were multiply aligned with the MUSCLE program, followed by tree reconstruction with the MEGA 7 software by using the JTT model and bootstrap value of 100 [22].

**PCR for the histone gene**

To avoid any possibilities of the misassembly of the histone gene in the assembled viral genome, a pair of PCR primers (5′-CTTCTGATATGGCCACCTTTGG-3′ and 5′-TCGCTTGGCGTCTTTCAAGG-3′) respectively targeting the up- and downstream of the histone gene were designed based on the assembled viral genomic sequence. PCR was then performed with the same DNA samples as applied to the metagenomic sequencing. The reaction (25 μL) contained 0.4 mM of forward and reverse primers, 12.5 μL of Taq PCR master mix (2×) (Tiangen Biotech), and 20–25 ng of DNA. The amplification started with the program: initial denaturation at 94 °C for 4 min, followed by 30 cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 45 s in an Eppendorf Mastercycler machine. PCR products were cloned and sequenced as previously described [15]. Sequences were aligned to the assembled genome by using Geneious R9 (Biomatters, https://www.geneious.com).

**Homology modeling of Histone H3 in DSLPV1**

The three dimensional structure of histone fold H3 in DSLPV1_013 was modeled on the online protein homology modeling SWISS-MODEL server (www.swissmodel.expasy.org) with the template of the crystal structure of human histone H3 (PDB id 3lel.1.E, amino acid sequence identity 83.82%). The structure was visualized and analyzed with the PyMOL software (www.pymol.org) [23].
**NCLDV conserved genes and virophage related genes in DSLPV1**

The nucleo-cytoplasmic virus orthologous group proteins (NCVOG) [24] were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG) for establishing a local database by using 'make blastdb' command in a BioLinux server. Similarly, all predicted proteins of Dishui Lake Virophage 1 (DSLV1, accession number KT894027) were downloaded to construct another local dataset. Predicted proteins of DSLPV1 were used as queries to search these two datasets respectively by using BLASTP program with cutoff e-value of 1e-3. The output results were recorded and analyzed for the NCLDV conserved genes and virophage related genes in DSLPV1.

**Nucleotide sequence accession number**

The DSLPV1 sequence has been deposited in the GenBank database (accession no. KY747489).

**Results**

**General organization of the DSLPV1 genome**

After the reference assembly, more than 8000 reads were eventually mapped to the consensus sequence (the contig that was initially confirmed to be related to algal large viruses, which is described in “Metagenomic sequence assembly of viral genome” in Methods.) with a high coverage (average coverage = 78.7, Additional file 1: Table S2) across the whole genome (Additional file 1: Figure S1), which indicated the accuracy of sequence assembly. The obtained genomic sequence with a size of 181,035 bp was named that of Dishui Lake Phycodnavirus 1 (DSLPV1). A 402-bp terminal inverted repeat was identified at both ends of the genome (Fig. 1a), which indicates the complete and linear genome of the DSLPV1. The G + C content of the genome is 52.7%, which is much higher than that of prasinoviruses (37.0–44.6%) but resembles YSLPVs (47.7–55.0%). A total of 227 ORFs were predicted, which are evenly distributed on both positive and negative DNA strands through the genome (i.e. 48% on the positive strand) with an average gene length of 770 bp and a coding density of 1.254 genes per kbp.

As the results of BLASTP search against NCBI nr database showed (Additional file 1: Table S3), among these 227 ORFs, 167 (~74%) shared similarity with proteins in nr database, with 155 having the top hits to algae large/giant viruses in NCLDV, 6 to bacteria, 4 to eukaryotes and 2 to archaea (Fig. 1 and Additional file 1: Table S3). Among the remaining 60 ORFs (~26%) having no hits in NCBI nr database (Fig. 1a), 35 had hits to NCBI environmental database (env Nr). Interestingly, 98% (152 of 155) ORFs having the best matches to giant/large viruses were homologous to algal large viruses in the Prasinovirus genus of the Phycodnaviridae family (Fig. 1b). Meanwhile, 75 out of these 155 ORFs with virus-hits showed no BLAST hits to cellular life (Additional file 1: Table S3), which are considered as virus specific genes, and the rest of 80 ORFs had both virus and non-virus hits (Additional file 1: Table S3). Of the four eukaryote-derived ORFs, one showed the best match to Ostreococcus lucimarinus CCE9901, the natural host of prasinoviruses, which indicates the potential genetic links between algae host and DSLPV1. In addition, two regions encoding tRNA (one Gln-tRNA and one Asn-tRNA) were identified in the DSLPV1 genome.

**Gene annotation in the DSLPV1 genome**

Functional annotation analysis of 227 putative proteins in DSLPV1 indicated that 160 (70.5%) had no homologues with defined functions in public databases. The remaining 67 putative proteins (29.5%) with annotated functions were classified into nine functional categories (Table 1). Ten putative proteins are involved in DNA replication, recombination and repair, which imply the independence of viral DNA replication in host cell; three putative proteins are associated with nucleotide transport and metabolism, eight with transcription, eight with protein and lipid synthesis/modification; seven encode capsid proteins; three have other miscellaneous functions, and one possibly possesses a signaling function.

**NCLDV conserved genes in the DSLPV1 genome**

DSLPV1 contains a nearly complete set of conserved genes in NCLDVs (40/47), including core genes of DNA polymerase B family protein, A18 helicase, A32 virion packaging ATPase and seven copies of major capsid protein. Ten NCVOGs in prasinoviruses were present in DSLPV1, and other key feature genes in prasinoviruses were also detected in DSLPV1 (Table 2). Notably, although some core genes were undetected in DSLPV1, e.g., serine/threonine protein kinase, they were probably functionally displaced by other genes. For instance, DSLPV1 ORF27 has a protein kinase structure which could be the functional substitute of the serine/threonine protein kinase. Such varied genomic features distinguish DSLPV1 from the prasinoviruses.

**Genomic architecture of DSLPV1**

Whole genome of DSLPV1 was aligned with that of the prasinoviruses (OtV5, MpV1, BpV2) and YSLPV1–3, respectively (Fig. 2a). DSLPV1 revealed good sequence synteny with the prasinoviruses, but shared few homologous segments with the YSLPVs. In addition, the gene content analysis also showed that DSLPV1 shared more homologous genes with these three prasinoviruses (OtV5: 59%, MpV1: 58%, BpV2: 61%) than with YSLPV1 (42%). These results suggest the DSLPV1 genomic architecture is more analogous to the prasinoviruses than to the YSLPVs.
Phylogenetic affiliation of DSLPV1

The annotated DNA polymerase gene (DSLPV1_189) of DSLPV1 shared high amino acid similarity (52–60%) with that of other phycodnaviruses, especially the prasinoviruses and the YSLPVs. Phylogenetic analysis of this gene and its homologues in other giant/large viruses indicated that DSLPV1 formed a monophyletic group with Prasinovirus with robust bootstrap support, and both DSLPV1 and the prasinoviruses are the sister lineages to the YSLPVs (Fig. 2b).

Host like genes in the DSLPV1 genome

A total of 44 ORFs in the DSLPV1 genome showed amino acid similarity (coverage > 70%, identity > 25%) to
genes found in the algae from both freshwater and marine environments, especially 12 ORFs that share high amino acid similarity (30–87%) to genes in the Mamiellophyceae class (green algae), containing natural hosts of the prasinoviruses (Table 3). Meanwhile, DSLPV1_139, the putative ribonucleoside-diphosphate reductase small subunit gene (rnr2) that is often shared by the prasinoviruses and their algae hosts, showed high amino acid similarity (63%) to the rnr2 in Ostreococcus lucimarinus CCE9901 (Table 3). In addition, DSLPV1_041 encodes a putative NAD-dependent epimerase/dehydratase protein that appears to derive from Ostreococcus lucimarinus CCE9901 as well (Table 3, Additional file 1: Figure S2).

Table 1: DSLPV1 ORFs with annotated functions (Continued)

| Function and protein | ORF |
|----------------------|-----|
| Patatin phospholipase | 106 |
| Asparagine synthetase | 181 |
| Capsid | |
| Capsid protein 1 | 53 |
| Capsid protein 2 | 59 |
| Capsid protein 3 | 81 |
| Capsid protein 4 | 82 |
| Capsid protein 5 | 147 |
| Capsid protein 6 | 188 |
| Capsid protein 7 | 194 |
| Virus inclusion body | 105 |
| Rhodanese domain-containing protein | 217 |
| Fibronectin-binding protein | 8 |

ORF numbers refer to that of DSLPV1. Their GenBank annotations are shown in Additional file 1: Table S3

Complete eukaryotic histone fold H3 in the DSLPV1 genome

Interestingly, DSLPV1 harbored a histone like gene (DSLPV1_013), which had high amino acid similarity (74–88%) to the histone gene from a wide range of eukaryotes. To eliminate the possibility of metagenomic sequence misassembly, the histone gene of DSLPV1_013 was re-checked with the DSLPV1 specific PCR and sequencing analysis. Sequence alignment analysis revealed that mismatches of the histone gene were not observed between the PCR amplified sequence and the assembled genome, which confirms the presence of a histone like gene in DSLPV1. Protein fold recognition with the InterProScan software detected a complete histone fold H3 at the C terminal of this protein (Fig. 3a). Moreover, sequence comparison of representative histone H3 folds
from eukaryotes, selected based on BLASTP results, with that of DSLPV1 revealed high conservation between them (Fig. 3b), and the phylogenetic analysis indicates that the histone gene of DSLPV1 is closely related to that of eukaryotes but not viruses (Fig. 3c).

Meanwhile, the predicted structure of DSLPV1 histone fold H3 revealed a canonical eukaryotic histone fold (Fig. 3d). In addition, the G + C content of the DSLPV1 histone gene (57%) is incongruous with that of the DSLPV1 genome (52.7%) but coherent with that of some

| Table 2 Conserved genes present in DSLPV1 and the prasinoviruses |
|----------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Gene name            | Otv5   |MpV1    | BpV1    | BpV2    | DSLPV1             |
| NCLDV core gene      | DNA polymerase    | 1       | 1       | 1       | 1                   |
|                      | ATPase (DNA packaging) | 1       | 1       | 1       | 1                   |
|                      | A18 helicase      | 1       | 1       | 1       | 1                   |
|                      | Capsid protein    | 8       | 8       | 7       | 7                   |
|                      | Serine/Threonine protein kinase | 1       | 1       | 1       | 0                   |
| NCVOGs present in prasinoviruses | Poxvirus late transcription factor VLTF3 | 1       | 1       | 1       | 1                   |
|                      | DNA topoisomerase | 2       | 1       | 1       | 1                   |
|                      | Transcription factor IIS | 1       | 1       | 1       | 1                   |
|                      | Transcription factor IIIB | 1       | 1       | 1       | 1                   |
|                      | Ribonucleotide reductase small subunit | 1       | 1       | 1       | 1                   |
|                      | Ribonucleotide diphosphate reductase large subunit | 1       | 1       | 1       | 1                   |
|                      | mRNA capping enzyme | 2       | 2       | 2       | 2                   |
|                      | ATP dependant DNA ligase | 1       | 1       | 1       | 0                   |
|                      | Exonuclease (Yqaj) | 2       | 2       | 2       | 2                   |
|                      | dUTP pyrophosphatase | 1       | 1       | 1       | 0                   |
|                      | Thymidine kinase | 1       | 1       | 1       | 0                   |
|                      | Thymidylate S | 1       | 1       | 1       | 0                   |
|                      | Nudix hydrolase | 1       | 1       | 1       | 0                   |
|                      | RuvC, Holliday junction resolvases, Extended Pox_A22, Poxvirus A22 family | 2       | 2       | 1       | 1                   |
| Other prasinovirus common genes | PCNA | 1       | 1       | 1       | 1                   |
|                      | NTPase/helicase | 1       | 1       | 1       | 1                   |
|                      | RNase H              | 1       | 1       | 1       | 1                   |
|                      | RNase III             | 1       | 1       | 1       | 1                   |
|                      | Transcription activator /SW/SNF helicase | 1       | 1       | 1       | 0                   |
|                      | TATA-box binding protein (TBP) | 1       | 1       | 1       | 1                   |
|                      | Prolyl 4-hydroxylase | 1       | 1       | 1       | 1                   |
|                      | 33 kDa in vitro translation peptide | 2       | 3       | 2       | 2                   |
|                      | 2OG-Fe(II) oxygenase | 1       | 1       | 1       | 1                   |
|                      | Aspartyl/Asparaginyl beta-hydrolase | 1       | 1       | 1       | 1                   |
|                      | Ubiquitin hydrolase-like cystein peptidase | 1       | 1       | 1       | 1                   |
|                      | FtsH (3–4-24) metalloendopeptidase | 1       | 1       | 1       | 1                   |
|                      | ATP-dependant protease proteolytic subunit | 1       | 1       | 1       | 1                   |
|                      | Patatin-like phospholipase | 1       | 1       | 1       | 1                   |
|                      | Potassium channel protein, PhoH | 1       | 1       | 1       | 1                   |
|                      | ABC-1 domain protein | 1       | 1       | 1       | 1                   |
|                      | ATP/GTP binding site motif A AGB-1 | 1       | 1       | 1       | 1                   |
|                      | Viral A-type inclusion protein | 1       | 1       | 1       | 1                   |

The numbers in the table represent that of gene copies
Fig. 2 (See legend on next page.)
algal genomes [25, 26]. Taken together, these results indicate the possibility of a recent horizontal transfer of the histone gene from eukaryotic hosts, e.g., algae, to viruses.

Genetic links between DSLPV1 and DSLV1

Intriguingly, three homologous genes that are shared between DSLPV1 and Dishui Lake Virophage 1 (DSLV1) were detected based on BLAST when searching the local database that consists of all predicted DSLV1 proteins by using the baits of 227 ORFs of DSLPV1. Functional annotation of these three homologous proteins revealed that the DSLV1_02 harbored two transmembrane regions at N terminal, which were analogous to DSLPV1_019 (26.6% of amino acid similarity) and DSLPV1_141 (25.4% of amino acid similarity). Similar RING-type zinc finger domain was detected at C terminal of DSLV1_25 and DSLPV1_175 (29.4% of amino acid similarity), and this domain is probably involved in mediating protein-protein interactions. In addition, a collagen triple helix repeat domain at C terminal of DSLV1_09 showed 41.2% of amino acid identity to DSLPV1_201, which represents the highest sequence similarity in comparison with the similarity that was shared between the other homologies described above.

Discussion

In this study, the genomic sequence of a novel phycodnavirus, named DSLPV1, was obtained from the DSL metagenomic datasets. The DSLPV1 genome appears to be complete as it contains a pair of 402-bp terminal inverted repeats flanking both ends of the linear genome as well as a nearly complete set of conserved genes of NCVOGs. Both comparative genomic and phylogenetic analyses indicate that DSLPV1 is closely affiliated to the prasinoviruses since they are similar in genomic length, share the most number of homologous genes including almost complete set of the Prasinovirus conserved genes (27/32), and formed a monophyletic clade based on the DNA PoIB protein phylogeny. Accordingly, DSLPV1 currently represents the closest freshwater relative of the marine prasinoviruses.

Dishui Lake where DSLPV1 was discovered is a holomictic freshwater lake (0.8‰ of salt) that was artificially excavated on tidal flat in 2003. Freshwater from neighboring Dazhi River is diverted into Dishui Lake, which is regularly discharged to the East China Sea. Interestingly, DSLPV1 is closely affiliated with the Prasinovirus infecting marine algae in the Mamiellophyceae class rather than with the freshwater-originated Prasinovirus-related

### Table 3 Host like genes in DSLPV1

| ORF       | Putative function                  | Hit   | Identity (%) | Coverage (%) | E value   |
|-----------|------------------------------------|-------|--------------|--------------|-----------|
| DSLPV1_008| Fibronectin-binding protein        | OL    | 38           | 95           | 8.00E-11  |
| DSLPV1_013| Histone H3                         | OL    | 87           | 71           | 1.00E-78  |
| DSLPV1_041| NAD-dependent epimerase            | OL    | 48           | 97           | 3.00E-89  |
| DSLPV1_043| Glycosyltransferase               | OT    | 27           | 69           | 5.00E-09  |
| DSLPV1_076| N-myristoyltransferase            | MP    | 36           | 99           | 1.00E-48  |
| DSLPV1_096| PCNA                               | MP    | 30           | 97           | 3.00E-31  |
| DSLPV1_113| DNA adenine methyltransferase     | Ms    | 33           | 93           | 2.00E-08  |
| DSLPV1_139| Ribonucleotide reductase, ss       | OL    | 63           | 73           | 2.00E-135 |
| DSLPV1_159| mRNA-capping enzyme                | OT    | 26           | 93           | 5.00E-24  |
| DSLPV1_164| Unknown                            | MP    | 46           | 96           | 3.00E-18  |
| DSLPV1_178| Transcription factor SII          | OL    | 33           | 98           | 8.00E-12  |
| DSLPV1_195| DNA topoisomerase II              | OT    | 43           | 100          | 0         |

OL: Ostreococcus lucimarinus CCE9901, OT: Ostreococcus tauri, MP: Micromonas pusilla CCMP1545, Ms: Micromonas sp. RCC299
Fig. 3 (See legend on next page.)

a) Diagram showing the histone fold/H3 and histone fold/H4 in various species. The species include:
- DSPLV1
- Ostreococcus tauri
- Aquila chrysaetos canadensis
- Lausannevirus
- Human adenovirus 50
- Influenza a virus

b) Amino acid sequences for the αN helix, α1 helix, and α2 helix of various species:
- Aquila chrysaetos canadensis
- Homo sapiens
- Noccaea caerulescens
- Zea mays
- Chlorella variabilis
- Bathycoccus prasioides
- Micromonas pusilla CCMP1545
- Micromonas commoda
- Ostreococcus tauri
- Ostreococcus lucimarinus CCE9901
- DSPLV1

For α3 helix:
- Aquila chrysaetos canadensis
- Homo sapiens
- Noccaea caerulescens
- Zea mays
- Chlorella variabilis
- Bathycoccus prasioides
- Micromonas pusilla CCMP1545
- Micromonas commoda
- Ostreococcus tauri
- Ostreococcus lucimarinus CCE9901
- DSPLV1

The sequences are shown with amino acid residues.

c) Flowchart showing the relationship between various species and the inclusion of:
- Entococcus silicicola
- Noccaea caerulescens
- DSPLV1
- Phaeodactylum tricornutum CCAP 1055/1
- Fragilaropsis cylindrus CCMP1102
- Thalassiosira pseudonana CCMP1335
- Ostreococcus lucimarinus CCE9901
- Micromonas commoda
- Ostreococcus tauri
- Rhooridium flavidum
- Micromonas pusilla CCMP1545
- Guillardia theta CCMP2712
- Bathycoccus prasioides
- Chondrus crispus
- Fucus serratus CAB82768.1
- Coccomyxa subellipsoidea C-169
- Chlorella variabilis
- Polyplethamides amyilfera
- Guldania sulphuraria
- Monoraphidium neglectum
- Marseillieviruses

Legend for the diagram:
- N terminal
- C terminal

Asterisk indicates a predicted or hypothetical sequence.
YSLPVs as well as with the members in the Chlorovirus genus that are presently the only well-defined freshwater algae viruses in Phycodnaviridae [27]. Viruses that are closely related to the prasinoviruses may have the Mamiellophyceae algae like host since all known prasinoviruses infect the Mamiellophyceae algae [28–30]. Meanwhile, some evidence showed the existence of freshwater algae species in Mamiellophyceae, albeit no phycodnaviruses were detected [10]. We consequently speculate that the host of DSLPV1 may be some kinds of freshwater Mamiellophyceae algae related species. Furthermore, the closest homologues of several viral genes in DSLPV1, especially the histone H3 gene, are present in the Ostreococcus algae as well, e.g., O. lucimarinus, O. tauri (Table 3). These genetic links between DSLPV1 and green algae not only further support the above proposed hypothesis, but also give better insight into a theoretical guidance for the isolation of potential algal host of DSLPV1. In addition, the Mamiellophyceae related contigs are the most abundant in the algae-related contigs that were identified in the DSL metagenomic datasets (data not shown). Hence, all these evidence indicates the possibility of the interaction relationship between DSLPV1 and the Ostreococcus related algae in freshwater. To further shed insight on the host of DSLPV1, both the DSLPV1 and its algae host need to be isolated, and infection experiments had to be performed. Notably, on the phylogenetic tree of the histone fold H3 protein (Fig. 3c), DSLPV1 was grouped together with the Noccaea caerulescens plant but not with the Ostreococcus algae. However, given that the bootstrap value is less than 50%, at almost all the branching points in the sub-tree comprising DSLPV1, the Ostreococcus algae and other representative eukaryotes, due to the high conservation of the histone fold H3 proteins, the phylogenetic affiliation of the DSLPV1 histone fold H3 protein remains uncertain. Accordingly, the horizontal transfer of the histone H3 fold protein gene may occur between DSLPV1 and the Ostreococcus algae. It hints again the Ostreococcus related algae might be the host of DSLPV1.

Surprisingly, a eukaryote derived histone like gene was identified in the genome of DSLPV1, which is firstly reported in the phycodnaviruses. It has been suggested that histones that associate to form the nucleosome and wrap the DNA in eukaryotic cells were probably acquired or mocked by viruses [31, 32]. Viral histones carry largely unknown functions [31–38]. Significant roles that are played by a few viral histones were demonstrated. The histone H4 protein in insect baculoviruses that shares high sequence identity with its host gene plays a critical role in suppressing host immune responses during infection by competing with endogenous cellular H4 for incorporation into the chromatin [39]. Interestingly, the nonstructural protein 1 (NS1 protein) that is encoded by some influenza virus strains, mimicry of the histone H3, suppresses the antiviral response of host by sequestering crucial factors required for transcriptional elongation [31]. Similarly, human adenovirus protein VII mimics cellular histone H3 for binding host nucleosomes to sequester immune danger signals to evade immune system during infection [40]. Alike the histone H4 in baculoviruses, the DSLPV1 histone H3 showed high amino acid similarity (87%) to the histone gene in its potential algae host except for an extra 53-aa tail at N terminal. In addition, sharing the same modification sites in N-terminal tail and canonical helices in core fold between DSLPV1 and the cellular histone H3 renders us to speculate the similar functions of the DSLPV1 histone. Regrettfully, the host of DSLPV1 is still unknown which debilitates the study of virus-host interaction to figure out the roles of viral histone in DSLPV1.

In our previous study, the genomes of algal large novel viruses YSLPVs and Yellowstone Lake giant virus (YSLGV) were discovered in Yellowstone Lake [14]. Meanwhile, interestingly the genomes of virophage YSLVs that are considered as the giant viral parasites were found in the same lake [16, 41]. Additionally, the genomes of virophages (Organic Lake Virophage, OLV, and Qinghai Lake Virophage, QLV) and their potential giant viral hosts, e.g., mimiviruses or phycodnaviruses, were obtained in Antartic Organic Lake and Qinghai Lake, China [6, 42]. Coincidently, in addition to DSLPV1, a novel virophage DSLV1 that is closely related to Yellowstone Lake Virophage 3 (YSLV3) in Yellowstone Lake was discovered in Dishui Lake as well [15]. Interestingly, DSLPV1 and DSLV1 share the collagen triple helix repeat containing protein that was
suspected to be involved in protein-protein interaction in virophage Sputnik/host mamavirus and virophage OLV/potential host Organic Lake Phycodnavirus (OLPV) associations [6, 43]. Genetic links were also observed between the YSLPVs and the YSLVs. However, their potential associations await future exhaustively experimental study.

Conclusions
In conclusion, a complete linear genomic sequence of DSLPV1 is discovered based on sequence assembly of the metagenomic datasets from Dishui Lake in Shanghai, China. Comprehensive genomic and phylogenetic analyses reveal that DSLPV1 represents a novel viral species in freshwater aquatic ecosystem and is closely related to the marine algae infecting Prasinovirus in the family Phycodnaviridae. Recent horizontal gene transfer was detected between DSLPV1 and its potential algal host. Our results here and the previous [14] suggest that the diversity and distribution of freshwater algal large or giant viruses remain far beyond exploration. Such knowledge will significantly contribute to better understanding not only the evolution of the phycodnaviruses and other related giant viruses, e.g., mimiviruses, but also their parasite viruses of virophages.

Additional file

Additional file 1: Supplementary data. It contains all supplementary figures and tables. (DOCX 447 kb)

Abbreviations
dsDNA: Double-stranded DNA; DSL: Dishui Lake; DSLPV1: Dishui Lake Phycodnavirus 1; DSLVI: Dishui Lake Virophage 1; NGLDV: Nucleo-cytoplasmic large DNA viruses; NCMC: Nucleo-Cytoplasmic Virus Orthologous Group; nr database: Non-redundant database; NS1 protein: Non-Structural protein 1; OLPV: Organic Lake Phycodnavirus; OLV: Organic Lake Virophage; ORF: Open Reading Frame; PCR: Polymerase Chain Reaction; QLV: Qinghai Lake Virophage; rnr2: ribonucleoside-diphosphate reductase small subunits; YSLGV: Yellowstone Lake giant virus; YSLPVs: Yellowstone Lake Phycodnaviruses; YSLV: Yellowstone Lake Virophage

Acknowledgements
We would like to thank Chaowen Gong, Yongxin Yu, Hongming Wang and Linghao Hu for their kind help on the assembly methodology, phylogenetic analyses and image preparation. We would also like to thank Taigang Liu for supporting on BioLinux servers.

Funding
This work was supported by the National Natural Science Foundation of China (31570112, 41376135).

Availability of data and materials
The complete DSLPV1 genome sequence has been deposited in GenBank (accession no. KY747489). The phylogenetic trees (Figs. 2b and 3c and Additional file 1: Figure S2) have been deposited in TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S22094). The datasets used and/or analyzed in the current study are available from the corresponding author on reasonable request.

Additional file

Additional file 1: Supplementary data. It contains all supplementary figures and tables. (DOCX 447 kb)

Abbreviations
dsDNA: Double-stranded DNA; DSL: Dishui Lake; DSLPV1: Dishui Lake Phycodnavirus 1; DSLVI: Dishui Lake Virophage 1; NGLDV: Nucleo-cytoplasmic large DNA viruses; NCMC: Nucleo-Cytoplasmic Virus Orthologous Group; nr database: Non-redundant database; NS1 protein: Non-Structural protein 1; OLPV: Organic Lake Phycodnavirus; OLV: Organic Lake Virophage; ORF: Open Reading Frame; PCR: Polymerase Chain Reaction; QLV: Qinghai Lake Virophage; rnr2: ribonucleoside-diphosphate reductase small subunits; YSLGV: Yellowstone Lake giant virus; YSLPVs: Yellowstone Lake Phycodnaviruses; YSLV: Yellowstone Lake Virophage

Acknowledgements
We would like to thank Chaowen Gong, Yongxin Yu, Hongming Wang and Linghao Hu for their kind help on the assembly methodology, phylogenetic analyses and image preparation. We would also like to thank Taigang Liu for supporting on BioLinux servers.

Funding
This work was supported by the National Natural Science Foundation of China (31570112, 41376135).

Availability of data and materials
The complete DSLPV1 genome sequence has been deposited in GenBank (accession no. KY747489). The phylogenetic trees (Figs. 2b and 3c and Additional file 1: Figure S2) have been deposited in TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:S22094). The datasets used and/or analyzed in the current study are available from the corresponding author on reasonable request.

Authors’ contributions
YW conceived and designed this study. HC and WZ assembled the sequences and analyzed the data. HC performed the experiments. XL performed the sampling and sequence analysis. SY and YP analyzed the data and modified the manuscript critically. HC, WZ and YW wrote the manuscript. YW analyzed the data. All authors read and approved the final manuscript.

Ethics approval and consent to participate
The water samples were collected from Dishui Lake with no permissions since it is freely available to the citizens.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1. College of Food Science and Technology, Shanghai Ocean University, Shanghai, China. 2. Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China. 3. Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation, Ministry of Agriculture, Shanghai, China. 4. Institute of Biochemistry and Molecular Cell Biology, University of Göttingen, Göttingen, Germany. 5. Present address: Archea Center, Department of Biology, Copenhagen University, DK2000 Copenhagen, Denmark.

Received: 29 November 2017 Accepted: 3 January 2018
Published online: 13 January 2018

References
1. Wilson WH, Van Etten JL, Schroeder DC, Nagasaki K, Brussaard C, Bratbak G and Suttle C. Family Phycodnaviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, editors. Virus taxonomy: Ninth report of the International Committee on Taxonomy of Viruses. Amsterdam: Elsevier Academic Press; 2012. p. 249–62.
2. Colson P, DeLambarerie X, Fournous G, Rocault D. Reclassification of giant viruses comprising a fourth domain of life in the new order Megavirales. Intervirology. 2012;55(5):321–32.
3. Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 2006;117(1):156–84.
4. Bratbak G, Helgadó M, Norland S, Thøngstad TF. Viruses as partners in spring bloom microbial trophodynamics. Appl Environ Microbiol. 1995;61(3):1400–5.
5. Martinez JM, Schroeder DC, Larsen A, Bratbak G, Wilson WH. Molecular dynamics of Emiliania huxleyi and cooccurring viruses during two separate mesocosm studies. Appl Environ Microbiol. 2007;73(2):554–62.
6. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA. Virophage control of marine viruses: a peek at genetic diversity. Virus Res. 2006;117(1):119–32.
7. Wilson WH, Van Etten JL, Chen et al. BMC Genomics (2018) 19:49 Page 11 of 12
