Distribution of soil viruses across China and their potential role in phosphorous metabolism

Li-Li Han (llhan@rcees.ac.cn)
Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences https://orcid.org/0000-0002-8105-1672

Dan-Ting Yu
Fujian Normal University

Li Bi
Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences

Shuai Du
Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences

Cynthia Silveira
San Diego State University

Ana Georgina Cobián Güemes
San Diego State University

Li-Mei Zhang
Research Centre for Eco-Environmental Sciences Chinese Academy of Sciences

Ji-Zheng He
Fujian Normal University

Forest Rohwer
San Diego State University

Research

Keywords: Virus, Virome, biogeography, latitude, phoH, P metabolism

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-361706/v1

Version of Record: A version of this preprint was published at Environmental Microbiome on February 7th, 2022. See the published version at https://doi.org/10.1186/s40793-022-00401-9.

Abstract

Background: Viruses are the most abundant biological entities on the planet and drive biogeochemical cycling on a global scale. Our understanding of biogeography of soil viruses and their ecological functions lags significantly behind that of Bacteria and Fungi. Here, a viromic approach was used to investigate the distribution and ecological functions of viruses from 19 soils across China.

Results: More than 60% of viral genome fragments could not be classified, representing potential new viruses. Among the 27 viral families identified, 15 families belonged to dsDNA viruses and 12 families belonged to ssDNA viruses. Soil samples clustered more significantly by geographical location than type of soil (agricultural and natural). Three clusters of viral communities were identified from North, Southeast and Southwest regions; these clusters differentiated using taxonomic as well as functional composition and were mainly driven by latitude. Phylogenetic analyses of the *phoH* gene showed a remarkable diversity and two new viral clades. Notably, five proteins involved in phosphorus (P) metabolism-related nucleotide synthesis functions, including dUTPase, MazG, PhoH, Thy1, and RNR, were mainly identified in agricultural soils.

Conclusions: The present work revealed that soil viral communities and their functions were distributed across China according to geographical location, with latitude as the dominant driving factor. In addition, P metabolism genes encoded by these viruses probably drive the synthesis of nucleotides for their own genomes inside bacterial hosts, thereby affecting P cycling in the soil ecosystems.

Introduction

Viruses are the most abundant and diverse biological life form and are major contributors to ecosystem functioning across all habitats [1]. Previous studies showed that viruses shape marine ecosystems by controlling the abundance and genomic diversity of their hosts through cell lysis [2-4] or lysogeny [5], and horizontal gene transfer [6-9]. Compared to around 1.01×10^{29} virus-like particles (VLPs) in marine environments, approximately 4.88×10^{30} VLPs were estimated to reside in global soils, accounting for 10% of the global viral abundance (4.80×10^{31}) [1]. The potential roles of soil viruses in terrestrial ecosystem processes include impacting microbial mortality, biogeochemical cycling of soil elements, and influencing food web dynamics [10]. Although soil viromes only contribute less than 1% of publicly available viral metagenomes [1], an increasing number of studies of viromes has focused on various soils, such as desert soil [11, 12], glacier soil [13], thawing permafrost soil [14], mangrove soil [9], mud volcanic soil [15], and Antarctic soil [16]. These studies revealed different patterns of soil viral community structure and largely uncharacterized viral assemblages. However, only a few studies have offered insight into how environmental factors influence viral communities. Soil pH, calcium content, and site altitude were the main drivers of the Antarctic viral community structure [16], and host community composition, pH, soil moisture content, and soil depth were correlated with viral communities along a thawing permafrost peatland soil [14].

In Chinese agricultural ecosystems, phosphorus (P) is an important biologically limiting nutrient that must be heavily supplemented for improving crop production [17]. Though lots of chemical P fertilizers have been applied to agricultural land [18], the P availability is still very low due to P slow diffusion and high fixation in soils. Previous studies showed that P content in the ecosystem could affect the proportion of P allocated from hosts to viruses, as viruses have a higher proportion of P (C/N/P≈20/6/1) [19] than Bacteria (69/16/1) [20, 21]. We considered the possibility that viruses accelerate the uptake of soil P to synthesize their own genomes when P fertilizers were supplemented by the host cell. Thus, viral infection could cause the P present in the host bacteria to be disproportionately incorporated into the new phage particles, further resulting in P removal from soil biotic cycling and affecting plant and microbial P acquisition strategies [22]. However, it is not clear how viruses manipulate this process and whether this process is related to the P concentration or P fertilizer input in the soils.

Increasing evidence has shown that abundant auxiliary metabolic genes (AMGs) encoded by viruses are expressed during the infection cycle, and that AMG products reprogram host cell metabolism with direct impacts on biogeochemistry [7, 23, 24]. In the genomes of globally abundant ocean viruses, more than two hundred viral-encoded AMGs have been identified [8], including carbon, nitrogen, sulfur, and P cycle related genes. However, as far as we know, only four types of viral AMGs (*trzN* [25], *phoH* [11], RNR [11] and CAZyme [9, 26]) have been identified in terrestrial ecosystems. Among them, only the *phoH* gene is presumed to belong to the Pho regulon and to regulate P uptake and metabolism under low-phosphate conditions [27]. Despite the ecological importance of viruses in soil biogeochemical cycling, our knowledge of their functions and potential mechanisms are notably limited.

In this study, we aimed to investigate the distribution of viral communities and functions from 19 soil samples across China, and determine the main factors driving viral distribution and function. Furthermore, we explored whether the *phoH* gene and its homologs may play important roles in P cycling in soil ecosystems.

Materials And Methods

2.1 Soil sampling and physicochemical properties

Between August 2015 and August 2016, a total of 19 soil samples were collected from 10 provinces across China; these samples included 10 agricultural soil samples and 9 natural soil samples (Fig. S1 and Table S1). The agricultural soil samples, from 5 maize fields and 5 paddy fields, were located in 7 provinces. The natural soil samples were also located in 7 provinces and included forest, grassland, wetland, coastal, glacier, and mud volcanic soils (Table S1). To study viral diversity and function in these soils, approximately 5 kg of each sample was collected and transported at 4°C back to the laboratory. At each site, a soil sample was collected from each of three separated 10 m × 10 m plots by pooling five upper 20-cm soil cores randomly taken from every plot. The three samples from each site were pooled and then processed as follows: 1 kg of soil was sieved to 1 mm for virus extraction, and 500 g of each soil sample was sieved to 2 mm and then stored at 4°C for physicochemical analyses.

A pH meter (Professional Meter PP-20, Sartorius, Germany) was used to measure soil pH and electrical conductivity (EC) at a ratio of 1:2.5 and 1:5 (soil to water, w/w), respectively. Organic matter (OM) was determined using the $K_2Cr_2O_7$ oxidation method, total nitrogen (TN) was measured using a Vario EL III analyzer (Elementar Analysensysteme GmbH, Hanau, Germany), available P was determined using the Olsen method [28], and available potassium (AK) was extracted with 0.5 M ammonium acetate and quantified using an atomic absorption spectrophotometer (ZEEnit700P, Analytik Jena AG, Jena, Germany). Mean annual temperature (MAT) and mean annual precipitation (MAP) data were from WorldClim Version2.

2.2 Virus extraction and purification

Viruses were extracted from the soil samples according to the method of Williamson et al. [29]. Briefly, 250 g of soil per sample was suspended in 1.5 L of glycine buffer (250 mM; pH = 8.5), shaken for 30 min, and centrifuged at 4000 g for 10 min at 4°C to precipitate soil particles. The supernatant was filtered sequentially through 1-mm, 0.45-μm, and 0.20-μm tangential flow filters (GE Healthcare Life Sciences, Pittsburgh, PA, USA). The viruses in the filtrate were concentrated using 30-kDa centrifugal ultrafiltration tubes (Merck Millipore Ltd., Tullagreen, Ireland) until the final sample volume was less than 1 ml. Finally, viral concentrates were treated with DNaseI (10 units DNaseI/100 μl) and incubated at 37°C for 1 h to remove free, non-encapsulated DNA. The presence of free and contaminating bacterial DNA was checked by PCR amplification of the 16S rRNA gene with primers 27F/1492R [30].

2.3 Viral DNA extraction and high-throughput sequencing

The Power Viral Environmental RNA/DNA Isolation kit (MO BIO Laboratories, Carlsbad, CA, USA) was used to extract total DNA. The REPLI-g Mini Kit (for multiple displacement amplification (MDA)) (Qiagen, Hilden, Germany) using Phi29 polymerase was applied to obtain the concentration and quantity needed for high-throughput sequencing. For each sample, 1 ng of DNA was fragmented to approximately 400 bp and used as a template to create a metagenome library, which was constructed according to the TruSeq™ DNA Sample Prep Kit (Illumina, San Diego, CA, USA) protocol. The libraries were loaded onto flow cell channels for sequencing using an Illumina HiSeq2500 at Shanghai Majorbio Bio-pharm Biotechnology Co., Ltd. (Shanghai, China) to generate 300-bp paired-end reads.

2.4 Analysis of viromes

2.4.1 Data sets and assembly

The original raw reads of the 19 samples obtained from the Illumina HiSeq2500 were cleaned using Fastq software for quality filtering and dereplication [31]. After quality control, each sample was independently assembled using metaSpades with default parameters, and contigs shorter than 1000 bp were eliminated [32]. Cloning vectors were removed using SMALT with 80% identity against the NCBI

UniVec database (until August, 2018). After removing the host contigs identified by CheckV v0.0.1 [33], all contigs were combined and clustered at 98% identity with cd-hit-est software [34], resulting in 64,579 non-redundant genome fragments used to create a viral Operational Taxonomic Units (vOTUs) database. Frap [1] was used to map clean reads from each sample to the vOTUs database at 90% identity, with the genome size normalization option, to obtain the normalized vOTU table. The number of viral reads was calculated by reads hitting these vOTUs.

2.4.2 Viral taxonomy clusters

An unsupervised random forest analysis was used to cluster the samples and identify which environmental and/or geographical factors influenced viral community composition and function using the "randomForest" and "rfPermute" packages on the R platform [35]. Each vOTU in the normalized vOTU table was assigned a value of 0 or 1 in order to create a vOTU presence-absence table to avoid MDA amplification bias, and to detect the viral contigs present in less than three soil samples. Non-metric multi-dimensional scaling (NMDS) was used to analyze the random forest proximity matrix, to cluster the samples based on Ward distances, and to identify the subset of variables of importance for the random forest clustering. The effects of environmental factors and geographical coordinates on this dataset were tested using a supervised random forest permutational-based variable importance measures to identify the significant predictors of viral community composition and function. UpSet analysis was further performed to visualize the interactive viral contigs among clusters by the "UpSetR" on the R platform [29].

2.4.3 Taxonomy annotation and comparison

A protein database (rcees_viral_database) was built from all viral sequences from the NR database (non-redundant proteins), the RefSeq viral protein database from the National Center for Biotechnology Information (NCBI), IMG/VR database and all phage databases from PHAST website (until August, 2018). The database was made non-redundant by clustering these genomes with cd-hit [34] with a 98% identity, resulting in a non-redundant dataset of 10,800,871 sequences. Each contig in the vOTU database was compared with the rcees_viral_database via BLASTx using Diamond, setting a minimum e value of 10^{-5} [36].

2.4.4 Functional analysis cluster

Open reading frames (ORFs) were predicted for viral contigs using metaProdigal [37, 38]. Predicted viral ORFs were compared to the Pfam domain database (version 32.0, release April, 2018) with a threshold of 40 for bit score and 0.001 for E-value [39]. The ORFs table was built the same way as the vOTU table. The ORF table was subjected to the same random forest, NMDS and UpSet analyses as was the vOTU table.

The contigs containing genes related to the P metabolism module were selected, including genes encoding the predicted proteins dUTPase [40], MazG, PhoH, Ribonucleoside triphosphate reductase (RNR), and Thymidylate synthase complementing protein (Thy1) [41, 42]. Genomic maps of contigs encoding the *phoH* gene were generated with Easyfig [43].

2.4.5 Phylogenetic analysis of the *phoH* gene

Phylogenetic trees of the *phoH* gene amino acid sequences were reconstructed using MEGA-X software [44]. A total of 773 reference sequences from viruses in paddy water and sea water were obtained from previous datasets [7, 27, 45]. All selected amino acid sequences were aligned by ClustalW, and the gaps and ambiguously aligned positions were deleted. After alignment, a phylogenetic tree was constructed using the Jones-Taylor-Thornton (JTT) model and the maximum likelihood method, and support for tree structure was obtained using 500 bootstrap pseudoreplicates.

2.5 Data availability

Virome read data are available in the NCBI Short Read Archive (SRA) under BioProject ID PRJNA579576.

Results

Viral community structure

Soil samples from 10 provinces across China were used to generate 19 soil viromes, including 10 agricultural soil viromes and 9 natural soil viromes (Fig. S1). A total of 186,503,518 reads (range: 4,782,132 to 15,945,343 per sample passed quality control. A total of 64,579 viral contigs (>1000 bp) were assembled; the longest contig was 98,359 bp and the average contig length was 2,288 bp. The de-replicated contigs formed the vOTU database. Contigs encoded between 2,395 and 40,474 ORFs (Table S2) and were used to identify 10,879,488 reads as viral according to a comparison with the vOTUs database.

To identify similarities between soils, soil viromes were clustered using an unsupervised random forest analysis. Three clusters of samples were identified (10.53% OOB estimate of error rate), and were related to the geographical distribution of the soil samples (Fig. 1a and Fig. S1). Cluster 1 included 7 of 8 North China samples, Cluster 2 included 4 of 5 Southeast China samples, and Cluster 3 included all 6 samples from the Southwest of China (Fig. 1a and Fig. S1). The 20 vOTUs with highest importance in differentiating clusters are shown in Fig. 1b and Fig. S2. The UpSet plot (Fig. 1c) showed that 88 viral contigs were shared by the three clusters, 437 contigs were shared by Clusters 1 and 2, 236 contigs were shared by Clusters 2 and 3, and 212 contigs were shared by Clusters 1 and 3. In addition, 618, 1138, and 2603 contigs were unique to Cluster 1, Cluster 2 and Cluster 3, respectively (Fig. 1c).

To analyze and compare the viral community composition with respect to environmental factors, Soil physical and chemical properties including pH, EC, OM, TN, AP, and AK, climate factors (MAT, Mean annual temperature, and MAP, mean annual precipitation (Table S1)), and geographical coordinates were tested as potential predictors of viral frequencies in the vOTU table. The results indicated that only MAT, longitude, and latitude explained 13.71%, 13.32%, and 26.71% of the variation in viral community

composition, respectively. Soil physical and chemical properties didn't show any relationship with viral community composition.

Viral genome completeness, estimated by CheckV, showed that 29% of the contigs in the vOTUs database were complete viral genomes (Fig. 2a). However, only 31% of contigs could be annotated by comparison with the rcees_viral_database and 62 % of contigs remained unclassified (Fig. 2b). A total of 27 viral families were identified from the 19 viromes, including 15 families belonging to dsDNA viruses and 12 families of ssDNA viruses. For ssDNA viruses, the *Microviridae* and *Circoviridae* families were widespread in all clusters, whereas *Geminiviridae*, *Cruciviridae*, and *Inoviridae* were mainly present in Cluster 1, *Bacilladnaviridae* and *Smacoviridae* in Cluster 2, and *Nanoviridae* in Clusters 2 and 3 (Fig. 2c). Among dsDNA viruses, more than half belonged to unclassified dsDNA viruses. The *Myoviridae* was widespread in all clusters, whereas *Siphoviridae*, *Podoviridae*, *Retroviridae*, and *Lavidaviridae* were mainly found in Clusters 1 and 2. The giant virus *Mimiviridae* was found in all three clusters but with relatively low abundances (Fig. 2d).

Clusters of viral-encoded functions

Three clusters of viral functional composition were identified by geographical distribution with 21.05% OOB estimate of error rate (Fig. 3a). Cluster 1 included 6 of 8 North China samples, Cluster 2 included 4 of 5 Southeast China samples, and Cluster 3 included all 6 Southwest China samples (Fig. 3a and Fig. S1). The 20 ORFs with highest importance to differentiate the functional clusters were mostly structural proteins (Fig. 3b and Fig. S3). The UpSet plot showed that 22 ORFs were shared by three clusters, 10 ORFs were shared by Clusters 1 and 2, 4 ORFs were shared by Clusters 2 and 3, and 104 ORFs were shared by Clusters 1 and 3. In addition, 22, 7, and 339 ORFs were unique to Cluster 1, Cluster 2 and Cluster 3, respectively (Fig. 3c).

Combined soil physical and chemical properties, climate factors (Table S1) and geographical coordinates explained a small percent of the variance in the functional profile frequencies, whereas latitude explained 15.47% variation in viral functional composition. However, soil physical and chemical properties didn't explain any relationship with viral functions.

P metabolism module

A phylogenetic tree of the *phoH* gene was built with 806 viral amino acid sequences from this study and others (Fig. 4). Sixteen representatives from 416 sequences were collected from Chinese paddy water [45], 5 representatives of 281 sequences from sea water [27], 5 representatives of 42 sequences from Global Ocean Survey metagenomes (GOS MGs) [7], 34 reference sequences from cultured phage, and 33 *phoH* amino acid sequences from this study. All of the 33 *phoH* amino acid sequences came from viromes of agricultural soils. Phylogenetic groups were defined according to first phylogenetic tree designed by Goldsmith *et al*. [27] and improved by Wang *et al*. [45]. Three *phoH* gene sequences from A6-

YAYM and 43 sequences from paddy water from China were grouped into Group δ. Out of the 33 sequences from this study, 23 sequences were separated into a new clade (group S1), which was completely different from other groups from marine and paddy water samples. Two *phoH* gene sequences from A6-YAYM and four *phoH* gene sequences from A3-DZ appeared in the group S2 with the cultured heterotrophic reference phages.

Genes functionally related to *phoH* were further analyzed using the PFAM domain database, which includes five P metabolism-related nucleotide synthesis functions involving dUTPase, MazG, PhoH, Thy1, and RNR. A total of 252 viral ORFs belonging to the five P metabolism proteins were identified (Fig. 5a), and they were mainly from agricultural soils (227 of 252 ORFs), especially from maize fields (209 of 252 ORFs). The representative contigs contained genes encoding dUTPase, Thy1 and RNR accompanied by the *phoH* gene in contigs A9_k141_8729, A9_k141_11614, A9_k141_52470, and A9_k141_80223 (Fig. 5b). In addition to these P metabolism proteins, these contigs encoded mostly hypothetical proteins.

Discussion

Soil plays a vital role in the distribution of organisms and their contributions to global biogeochemical cycles [22]. However, our understanding of soil viruses, and the factors driving their distribution, lags far behind that of marine viruses. Furthermore, the ways in which viruses participate in the biogeochemical cycling of soil elements has not been extensively investigated. This study provides evidence that latitude is a key driver of viral distribution in soils. Furthermore, the higher abundance of viral-encoded P metabolism genes in agricultural soils indicates that viruses have the potential roles of P cycling in these soil ecosystems.

The taxonomic distribution of soil viruses

The order *Caudovirales* (tailed viruses that infect Bacteria and Archaea), including *Siphoviridae*, *Myoviridae* and *Podoviridae*, was dominant in all of our soil samples, in agreement with previous studies [9, 11, 12, 16, 46]. In the Antarctic soil, *Podoviridae* presented at similar levels in all samples, whereas the abundances of *Myoviridae* and *Siphoviridae* were inversely correlated, as they may have direct competition for hosts in the same niche, and Siphoviruses are always present at higher abundances in neutral to alkaline pH soils [16]. However, our study showed different trends, with *Myoviridae* occurring in all samples, whereas *Siphoviridae* and *Podoviridae* were mainly present in more acidic soils. More data may be needed to find patterns, especially since so many viruses in the viromic data could not be classified.

Moreover, our soil viromes revealed diverse ssDNA viruses belonging to the *Microviridae*, *Circoviridae*, *Genomoviridae*, and *Cruciviridae* (Fig. 1c). The broad presence of ssDNA viruses is possibly due to the bias of MDA, which is necessary to generate enough viral DNA for metagenomic analyses but preferentially amplifies genomes of ssDNA viruses and thus leads to a quantitative bias [47]. Therefore, we analyzed the dsDNA and ssDNA viruses separately, and ssDNA viruses were reported in a qualitative rather than quantitative way in this study.

Latitude drives viral community composition and function

Viral community composition has been associated with a variety of environmental factors, such as host community composition, pH, soil depth and moisture [14], calcium content and site altitude [16]. According to the unsupervised random forests analysis, the viral communities and functions from 19 soil samples across China grouped into 3 clusters, which corresponded to geographical location well (Fig. 1a, Fig. 3a and Fig. S1). A subsequent supervised random forest analysis showed that the main environmental driver of these clusters for both viral communities and functions was latitude. There have been few reports regarding location and its effects on the distribution of viruses. Such as the altitude of Antarctic soils which probably linked to temperature could influence microbial metabolism and substantially impact viral communities and functions [16]. The temperature change along the latitude in this study may have similar effects, especially on viral community. All of the viruses differentiating these clusters were unclassified viruses. This highlighted the lack of knowledge and reference sequences for soil viruses.

Although phosphorus is an important factor of viral genome synthesis, the results do not imply any relationship between soil available P content and viral communities and functions. It is possible that our sampling time may be at different stages of phosphorus metabolism because of different fertilization time in each agricultural region. On the other hand, soil available P content may affect viral abundance more than viral community, and we will further focus on this point in the future.

Viruses may directly manipulate P cycling in soils

The *phoH* gene has been widely used as a signature gene for assessing viral phylogeny and diversity, and is encoded by various morphologically distinct viruses that infect a wide range of hosts, including autotrophic and heterotrophic Bacteria and Eukaryotes [27, 28]. A diversity of *phoH* genes have been found in viral communities inhabiting numerous environments, such as seawater [27], paddy water [45], and a Namib hypolith [11]. In these studies, *phoH* genes were distributed according to depth and location [27], biogeography [45], or were found to be entirely novel [11]. In this study, phylogenetic analyses showed that a few *phoH* sequences in groups 1, 2, 3 and δ (Fig. 4) were widely distributed in agricultural soils, paddy water [45] and sea water from different sites of the world [7, 27]. However, most viral *phoH* sequences (29/33) in this study belonged to two new viral clades (groups S1 and S2) that clearly differed from those in marine and paddy waters (Fig. 4). The majority of the Namib hypolith *phoH* amino acid sequences clustered separately from other sequences and was omitted from our phylogenetic tree. These results support the inference that the distribution of viral *phoH* genes is dependent on characteristics of the environment [48].

During the second Chinese soil survey [49], a database created from 2,473 soil profiles was analyzed and showed relatively consistent C:P (136) and N:P (9.3) ratios, with a highly constrained C:N:P ratio of 134:9:1 for the surface soils from both of agricultural and natural soils [50]. This ratio indicates that the P content in Chinese soils is generally lower than that required by phages, which have a C:N:P ratio of 20:6:1 [19]. Due to P slow diffusion and high fixation in soils, plus the crops on the absorption of P for

agricultural production [18], this means that P can be a major limiting factor for soil microbes, especially viruses. Based on this background, this P deficient environment may select for these viruses to regulate P uptake and metabolism through evolution of the *phoH* gene. It is interesting that all 33 *phoH* gene sequences identified in this study were from viruses in agricultural soils. It is possible that agricultural soil is a rich environment in terms of dissolved organic matter, produced via photosynthesis, and nitrogen applied as fertilizer, but that these excesses of C and N result in P being limited. Once P fertilizer input, virus may prompt its host to quickly absorb inorganic P (Pi) and use PhoH to help its own reproduction (Fig. 6).

To better understand the metabolic potential of *phoH* genes, we searched for, but did not find, additional genes in the Pho regulon. However, it is interesting that four functions related to nucleotide synthesis, including dUTPase, MazG, Thy1, and RNR, were identified in association with *phoH* to act as a P metabolism module. Previous studies have demonstrated the presence of at least five proteins involved in P metabolism including PhoH, RNR, Thy1, endodeoxyribonuclease, and MazG pyrophosphatase in marine phage genomes [41, 42]. Similar modules were also found in two complete viral genomes from two agricultural soils in our unpublished data, including dUTPase, PhoH, RNR, and Thy1 (Fig. S4). Until our study, only two P metabolism genes (*phoH* and RNR) have been reported in terrestrial ecosystems [11]. Here, five of the P metabolism genes were identified, especially in agricultural soils (Fig. 5a). Among them, MazG is reported as a nucleoside triphosphate pyrophosphohydrolase, which can hydrolyze all eight of the canonical ribo- and deoxynucleoside triphosphates to their respective monophosphates and PP(i), with a preference for deoxynucleotides [51]. RNR, known as ribonucleoside diphosphate reductase, converts all four ribonucleotide diphosphates (rNDPs) to the respective deoxynucleoside diphosphates (dNDPs), which are then rapidly converted to dNTP [41, 52]. The dUTPase can catalyze dUTP to dUMP and release diphosphate, and provide a substrate (dUMP) for thymidylate synthase [40]. Thy1 can convert dUMP to dTMP depending on FAD, NADPH and 5, 10-methylenetetrahydrofolate [53]. PhoH has been reported as a cytoplasmic protein with an ATP-binding activity and is predicted to be induced by P starvation [54]; however, its function remains unknown. Altogether, this information led us to hypothesize that PhoH can act as a nucleotide synthase, possibly binding and hydrolyzing ATP through its conserved nucleoside triphosphate hydrolase domain to obtain energy, and taking advantage of Pi from the agricultural soil (through the host cell) to catalyze the synthesis of nucleotides for the virus's own genome (see conceptual model in Fig. 6). This model predicts the proliferation of a huge number of soil viruses playing an important role in depleting P from the soil ecosystem. Future work should focus on whether the concentration of Pi in soil is associated with the number of progeny produced by viruses, and also quantify the contribution of viruses to P loss from soil.

Conclusions

In summary, our analyses systematically explored viral community structure and function in soils across China. The results revealed that the distribution of viral communities and their functions was at least partly determined by geographical location especially latitude. Remarkably, AMGs related to P metabolism, including PhoH, RNR, Thy1, dUTPase and MazG, were mainly identified in viral genomes

from agricultural soils, which suggested that viruses possibly take advantage of the Pi added to agricultural soils to synthesize their own genomes. As a consequence, these soil viruses have the potential to significantly contribute to P cycling in the soil ecosystem. Future investigations of the relationship between soil Pi content and viral ecology will reveal the specific mechanism of viral genome synthesis using soil-derived P and resulting depletion of soil P and provide more detailed insights into the contributions of viruses to the P cycle in soil ecosystems.

Declarations

Acknowledgements

We thank Dr. Taylor O'Connell, Mark Little, Nate Robinett, and Adam Barno at San Diego State University for their generous help with data analyses and paper writing.

Funding

This work was supported by the National Science Foundation of China (grant number: 41771289 and 41571248), the National Key R&D Program (grant number: 2017YFD0200600), and the China Scholarship Council.

Availability of data and materials

Virome read data are available in the NCBI Short Read Archive (SRA) under BioProject ID PRJNA579576.

Authors' contributions

LLH, LMZ and JZH designed the research. LLH, DTY, LB and SD sampled the soils and conducted the laboratory analyses and the raw data collection. LLH, JZH, CS, AGC and FR performed the data processes and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable as there were no human, animal or pathogen subjects involved.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F: **Viruses as winners in the game of life.** *Annual Review of Virology* 2016, **3:**197-214.
2. Thingstad TF: **Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems.** *Limnology and Oceanography* 2000, **45:**1320-1328.
3. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R: **Viral and microbial community dynamics in four aquatic environments.** *The ISME journal* 2010, **4:**739.
4. Thingstad TF, Våge S, Storesund JE, Sandaa R-A, Giske J: **A theoretical analysis of how strain-specific viruses can control microbial species diversity.** *Proceedings of the National Academy of Sciences* 2014, **111:**7813-7818.
5. Knowles B, Silveira C, Bailey B, Barott K, Cantu V, Cobián-Güemes A, Coutinho F, Dinsdale E, Felts B, Furby K: **Lytic to temperate switching of viral communities.** *Nature* 2016, **531:**466.
6. Rohwer F, Thurber RV: **Viruses manipulate the marine environment.** *Nature* 2009, **459:**207-212.
7. Sharon I, Battchikova N, Aro E-M, Giglione C, Meinnel T, Glaser F, Pinter RY, Breitbart M, Rohwer F, Béjà O: **Comparative metagenomics of microbial traits within oceanic viral communities.** *The ISME journal* 2011, **5:**1178.
8. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J: **Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses.** *Nature* 2016, **537:**689-693.
9. Jin M, Guo X, Zhang R, Qu W, Gao B, Zeng R: **Diversities and potential biogeochemical impacts of mangrove soil viruses.** *Microbiome* 2019, **7:**58.
10. Emerson JB: **Soil viruses: a new hope.** *MSystems* 2019, **4**.
11. Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA: **Metagenomic analysis of the viral community in N amib D esert hypoliths.** *Environmental Microbiology* 2015, **17:**480-495.
12. Scola V, Ramond J-B, Frossard A, Zablocki O, Adriaenssens EM, Johnson RM, Seely M, Cowan DA: **Namib desert soil microbial community diversity, assembly, and function along a natural xeric gradient.** *Microbial Ecology* 2018, **75:**193-203.
13. Han L-L, Yu D-T, Zhang L-M, Wang J-T, He J-Z: **Unique community structure of viruses in a glacier soil of the Tianshan Mountains, China.** *Journal of Soils and Sediments* 2017, **17:**852-860.
14. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA: **Host-linked soil viral ecology along a permafrost thaw gradient.** *Nature Microbiology* 2018, **3:**870.
15. Yu D-T, He J-Z, Zhang L-M, Han L-L: **Viral metagenomics analysis and eight novel viral genomes identified from the Dushanzi mud volcanic soil in Xinjiang, China.** *Journal of Soils and Sediments* 2019, **19:**81-90.

16. Adriaenssens EM, Kramer R, Van Goethem MW, Makhalanyane TP, Hogg I, Cowan DA: **Environmental drivers of viral community composition in Antarctic soils identified by viromics.** *Microbiome* 2017, **5:**83.
17. Kirkby EA, Johnston AEJ: **Soil and fertilizer phosphorus in relation to crop nutrition.** In *The ecophysiology of plant-phosphorus interactions.* Springer; 2008: 177-223
18. Qiu J: **Phosphate fertilizer warning for China.** *Nature News* 2010.
19. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz J: **The elemental composition of virus particles: implications for marine biogeochemical cycles.** *Nature Reviews Microbiology* 2014, **12:**519.
20. Suttle CA: **Marine viruses-major players in the global ecosystem.** *Nature Reviews Microbiology* 2007, **5:**801.
21. Sterner RW, Elser JJ: *Ecological stoichiometry: the biology of elements from molecules to the biosphere.* Princeton university press; 2002.
22. Kuzyakov Y, Mason-Jones K: **Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions.** *Soil Biology and Biochemistry* 2018, **127:**305-317.
23. Breitbart M, Thompson LR, Suttle CA, Sullivan MB: **Exploring the vast diversity of marine viruses.** *Oceanography* 2007, **20:**135-139.
24. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW: **Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism.** *Proceedings of the National Academy of Sciences of the United States* 2011, **108:**E757-E764.
25. Ghosh D, Roy K, Williamson KE, White DC, Wommack KE, Sublette KL, Radosevich M: **Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and trzN genes in viral-community DNA.** *Applied and Environmental Microbiology* 2008, **74:**495-502.
26. Bi L, Yu DT, Du S, Zhang LM, Zhang LY, Wu CF, Xiong C, Han LL, He JZ: **Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils.** *Environmental Microbiology* 2021, **23:**588-599.
27. Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa R-A, Breitbart M: **Development of phoH as a novel signature gene for assessing marine phage diversity.** *Applied and Environmental Microbiology* 2011, **77:**7730-7739.
28. Olsen SR: **Estimation of available phosphorus in soils by extraction with sodium bicarbonate.** *United States Department of Agriculture* 1954, **939:**1-19.
29. Williamson KE, Radosevich M, Wommack KE: **Abundance and diversity of viruses in six Delaware soils.** *Applied and Environmental Microbiology* 2005, **71:**3119-3125.
30. DeLong EF: **Archaea in coastal marine environments.** *Proceedings of the National Academy of Sciences of the United States* 1992, **89:**5685-5689.
31. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets.** *Bioinformatics* 2011, **27:**863-864.

32. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler.** *Genome Research* 2017, **27:**824-834.
33. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides N: **CheckV assesses the quality and completeness of metagenome-assembled viral genomes.** *Nature Biotechnology* 2020.
34. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22:**1658-1659.
35. Cutler A, Cutler DR, Stevens JR: **Random forests.** In *Ensemble machine learning.* Springer; 2012: 157-175
36. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND.** *Nature Methods* 2015, **12:**59.
37. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11:**119.
38. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC: **Gene and translation initiation site prediction in metagenomic sequences.** *Bioinformatics* 2012, **28:**2223-2230.
39. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J: **Pfam: the protein families database.** *Nucleic Acids Research* 2013, **42:**D222-D230.
40. Chen R, Wang H, Mansky LM: **Roles of uracil-DNA glycosylase and dUTPase in virus replication.** *Journal of General Virology* 2002, **83:**2339-2345.
41. Rohwer F, Segall A, Steward G, Seguritan V, Breitbart M, Wolven F, Azam F: **The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages.** *Limnology Oceanography* 2000, **45:**408-418.
42. Angly F, Youle M, Nosrat B, Srinagesh S, Rodriguez-Brito B, McNairnie P, Deyanat-Yazdi G, Breitbart M, Rohwer F: **Genomic analysis of multiple Roseophage SIO1 strains.** *Environmental Microbiology* 2009, **11:**2863-2873.
43. Sullivan MJ, Petty NK, Beatson SA: **Easyfig: a genome comparison visualizer.** *Bioinformatics* 2011, **27:**1009-1010.
44. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: molecular evolutionary genetics analysis across computing platforms.** *Molecular Biology and Evolution* 2018, **35:**1547-1549.
45. Wang X, Liu J, Yu Z, Jin J, Liu X, Wang G: **Novel groups and unique distribution of phage phoH genes in paddy waters in northeast China.** *Scientific Reports* 2016, **6:**38428.
46. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, Cowan D: **High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils.** *Applied and Environmental Microbiology* 2014, **80:**6888-6897.
47. Reavy B, Swanson MM, Cock PJ, Dawson L, Freitag TE, Singh BK, Torrance L, Mushegian AR, Taliansky M: **Distinct circular single-stranded DNA viruses exist in different soil types.** *Applied and Environmental Microbiology* 2015, **81:**3934-3945.

48. Xiang L, Yan S, Xin-zhen W, Jun-jie L, Guang-hua W: **Research Progress of New Biomarker Gene of phoH for Bacteriophage Genetic Diversity.** *Biotechnology Bulletin* 2017, **33:**40-45.
49. Shi X, Yu D, Warner E, Pan X, Petersen G, Gong Z, Weindorf D: **Soil database of 1: 1,000,000 digital soil survey and reference system of the Chinese genetic soil classification system.** *Soil Survey Horizons* 2004, **45:**129-136.
50. Tian H, Chen G, Zhang C, Melillo JM, Hall CA: **Pattern and variation of C: N: P ratios in China's soils: a synthesis of observational data.** *Biogeochemistry* 2010, **98:**139-151.
51. Zhang J, Inouye M: **MazG, a nucleoside triphosphate pyrophosphohydrolase, interacts with Era, an essential GTPase in Escherichia coli.** *Journal of Bacteriology* 2002, **184:**5323-5329.
52. Stillman B: **Deoxynucleoside triphosphate (dNTP) synthesis and destruction regulate the replication of both cell and virus genomes.** *Proceedings of the National Academy of Sciences* 2013, **110:**14120-14121.
53. Ogawa A, Sampei G-i, Kawai G: **Crystal structure of the flavin-dependent thymidylate synthase Thy1 from Thermus thermophilus with an extra C-terminal domain.** *Acta Crystallographica Section F: Structural Biology Communications* 2019, **75:**450-454.
54. Kim S, Makino K, Amemura M, Shinagawa H, Nakata A: **Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli.** *Journal of Bacteriology* 1993, **175:**1316-1324.

Figures

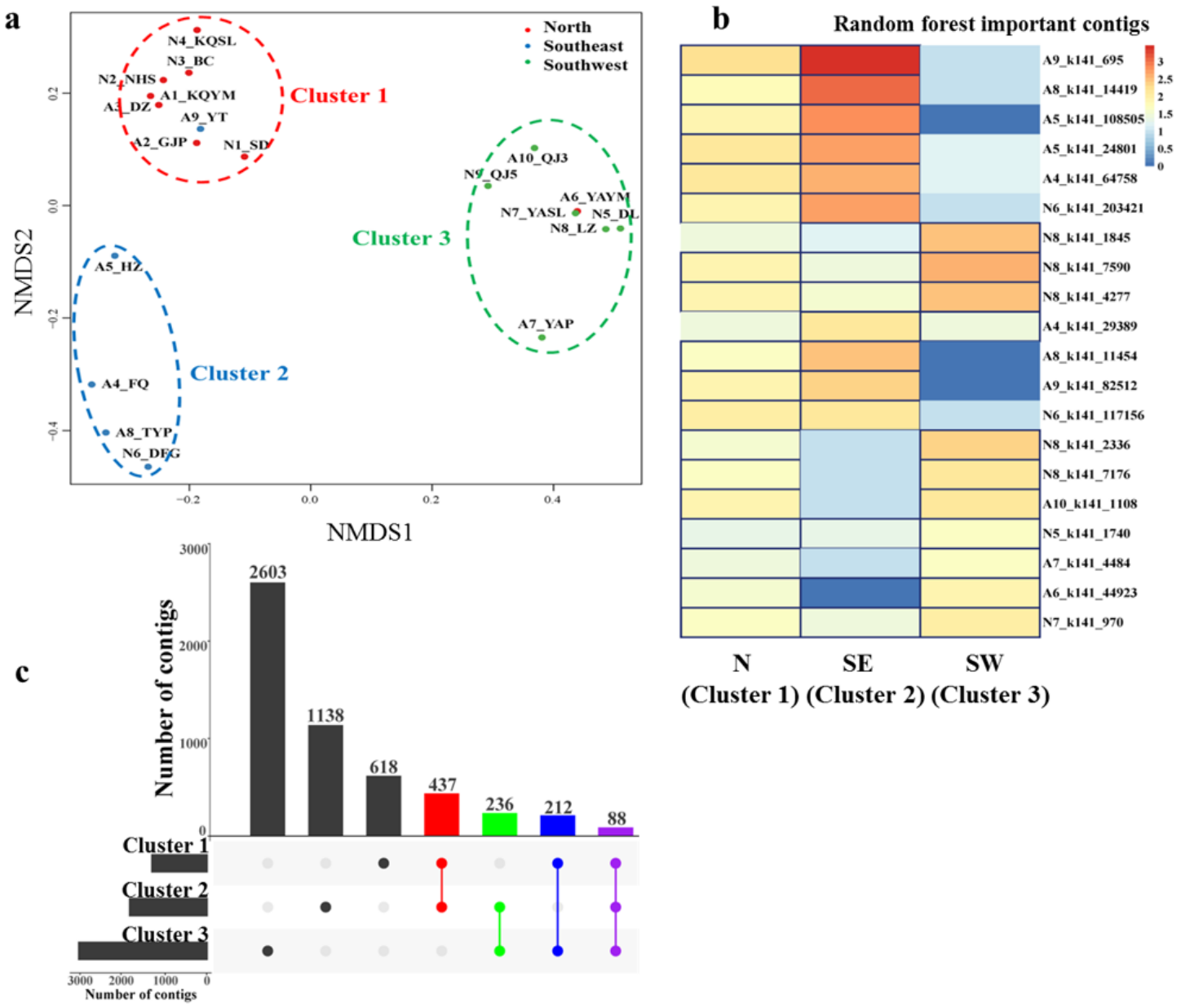

Figure 1

Viral community structure. a) Non-metric multidimensional scaling (NMDS) of viral community composition in 19 soil viromes obtained from an unsupervised random forest analysis followed by clustering using Ward distances (OOB estimate of error rate = 10.53 %). Symbols are color-coded by site (red: North sites; blue: Southeast sites; green: Southwest sites). b) Contigs differentiating the geographical clusters in a random forest analysis supervised by geographical location. The ranks surrounded by a black box indicate variables with p-values less than 0.01 in the permutational test. c) Shared viral contigs among three clusters displayed by UpSet analysis.

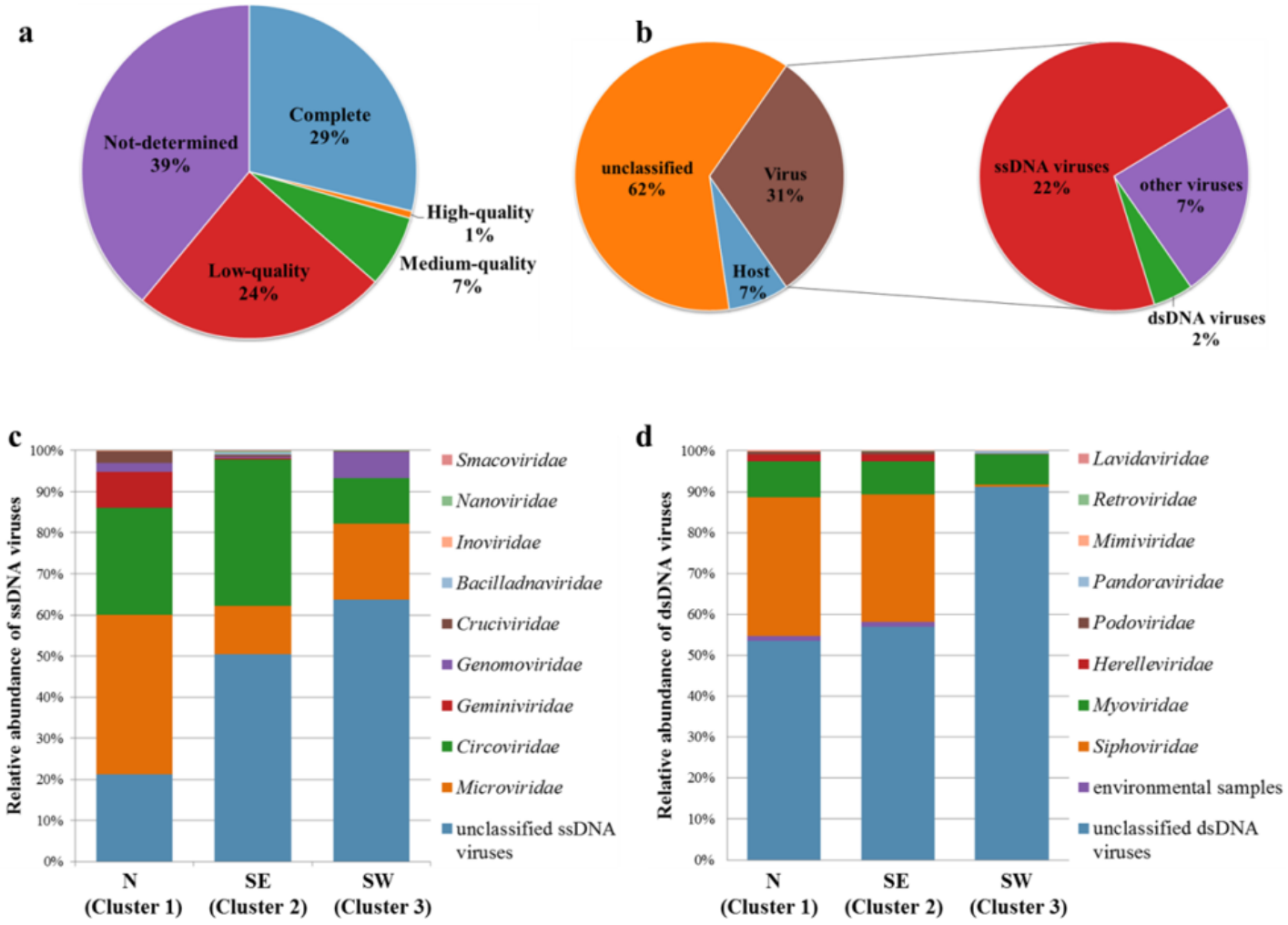

Figure 2

Viral taxonomy. a) Genome completeness in the full dataset as determined by CheckV (complete, high-quality (>90% completeness), medium-quality (50-90% completeness), low-quality (0-50% completeness), and undetermined-quality). b) Percentage of annotated viral sequences. c) Relative abundances of the dominant ssDNA viral families among three NMDS clusters (top 10 are shown). d) The relative abundances of the dominant dsDNA viral families among three NMDS clusters (top 10 are shown).

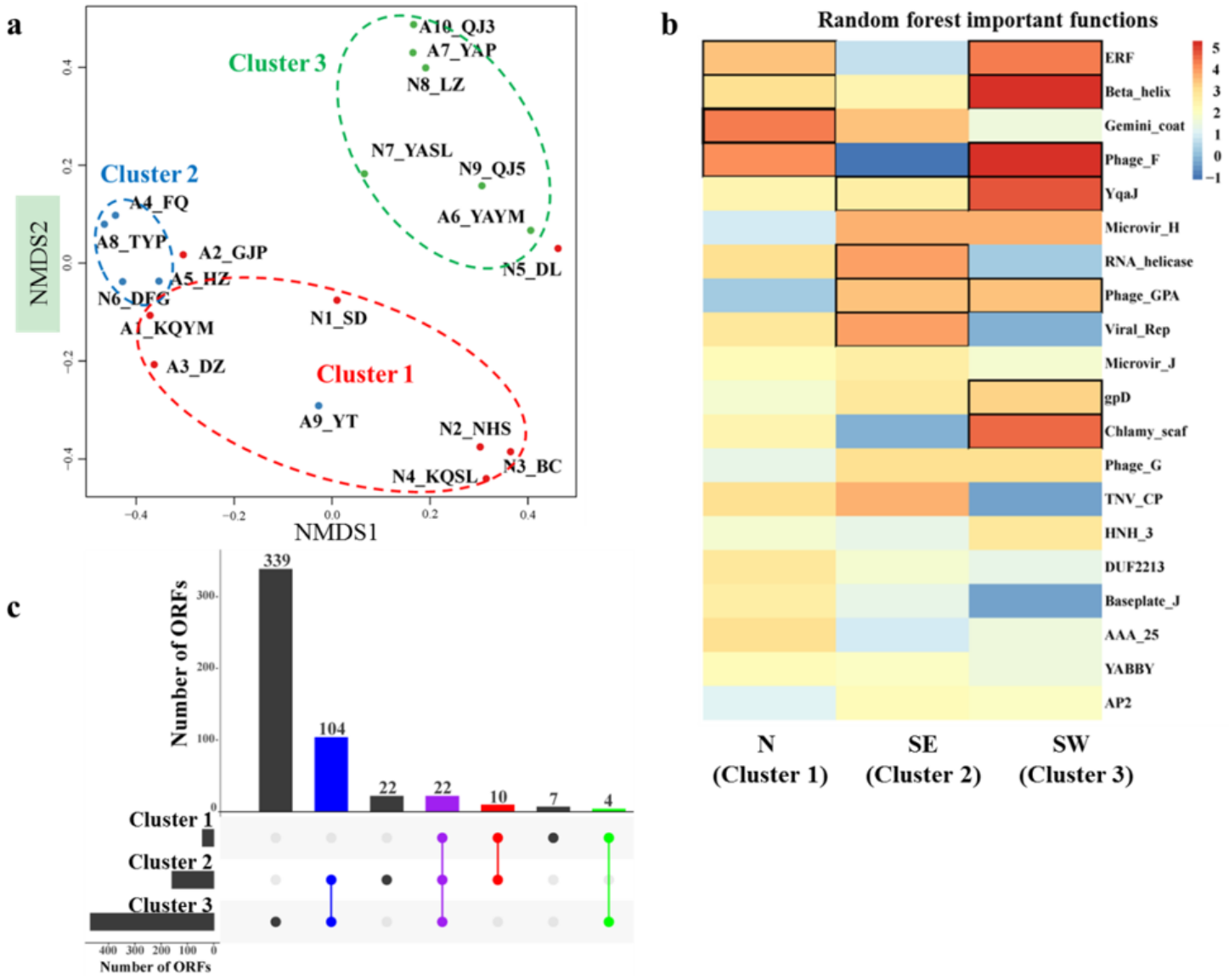

Figure 3

Viral functional clusters. a) NMDS of viral functions of 19 soil viromes by an unsupervised random forest analysis followed by clustering using Ward distances (OOB estimate of error rate = 26.32 %). Symbols are color-coded by site (red: North site; blue: Southeast site; green: Southwest site). b) Important functions differentiating the geographical distribution in a random forest analysis supervised by sites. The ranks surrounded by a black box indicate variables with p-values less than 0.01 in the permutational test. c) Shared viral ORFs among three clusters displayed by UpSet analysis.

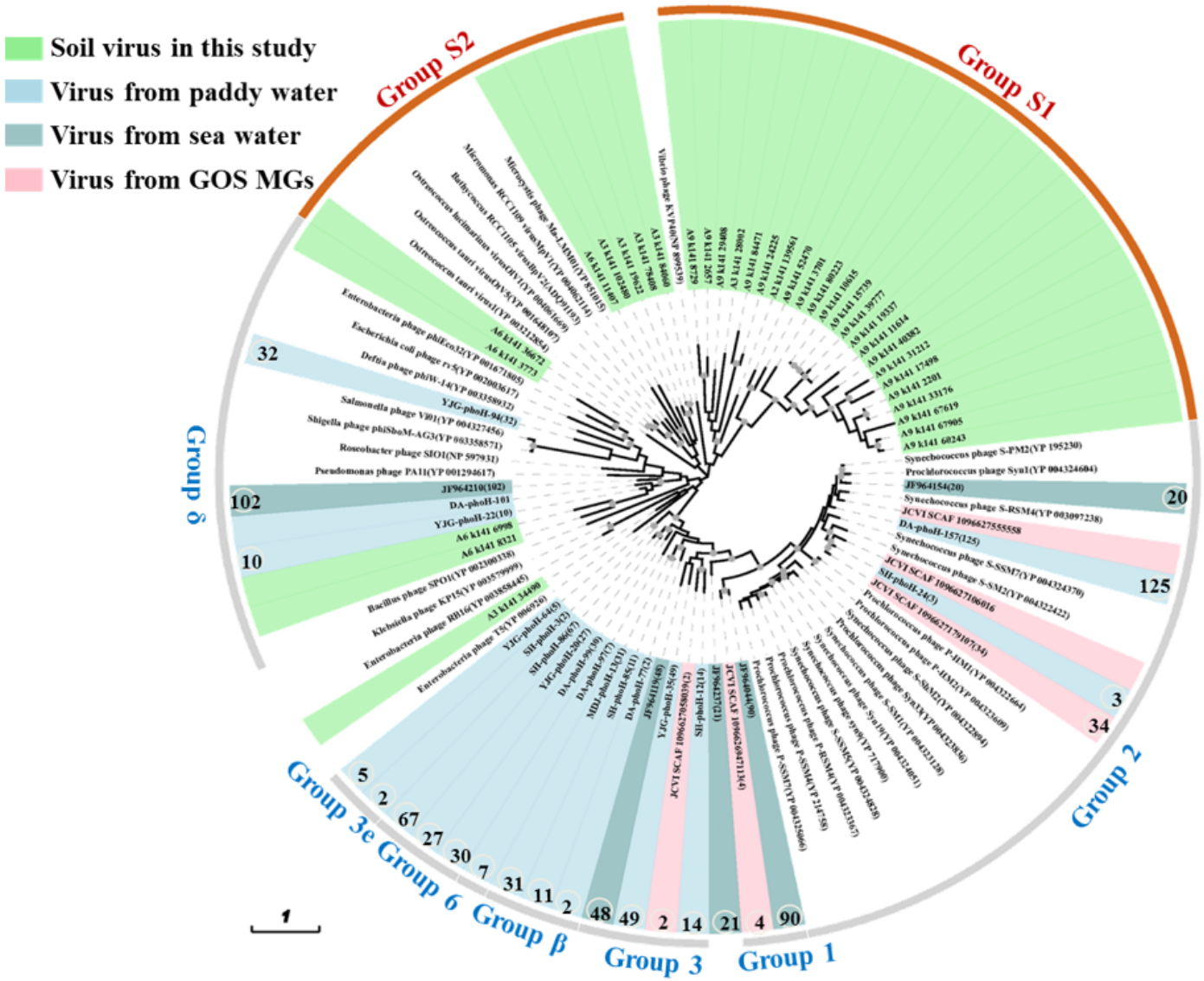

Figure 4

Phylogenetic tree of phoH gene sequences. The groups were defined according to first designation by Goldsmith et al. [27] and improved by Wang et al. [45]. Only representative sequences are displayed in the tree, and the number represents the number of original sequences.

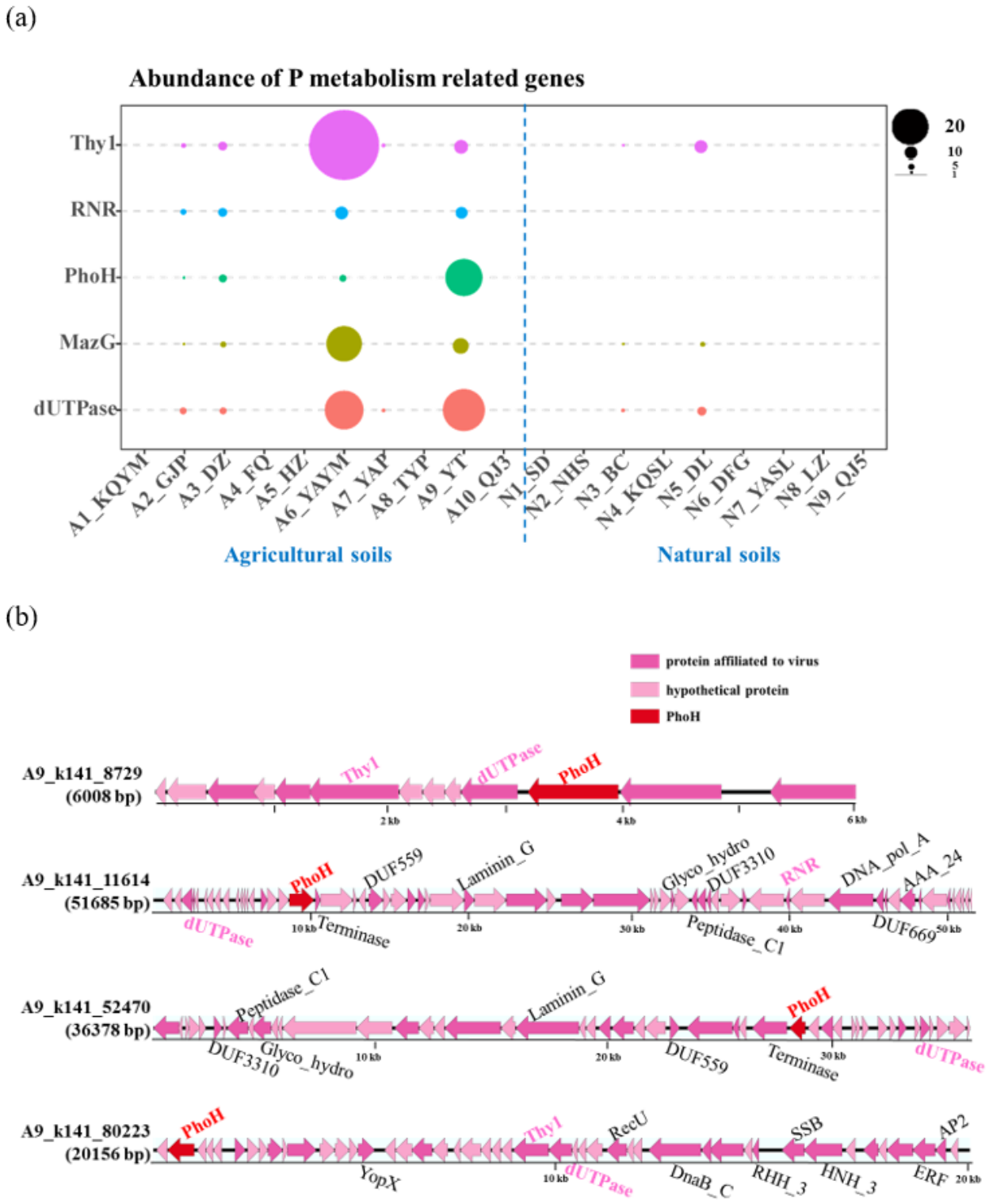

Figure 5

(a) The abundances of genes in the Phosphate metabolism module, shown for each site. Circle size represents the number of sequences detected. (b) Examples of viral contigs carrying AMGs involved in phosphate metabolism. Arrows indicate ORFs and arrow color indicates predicted function.

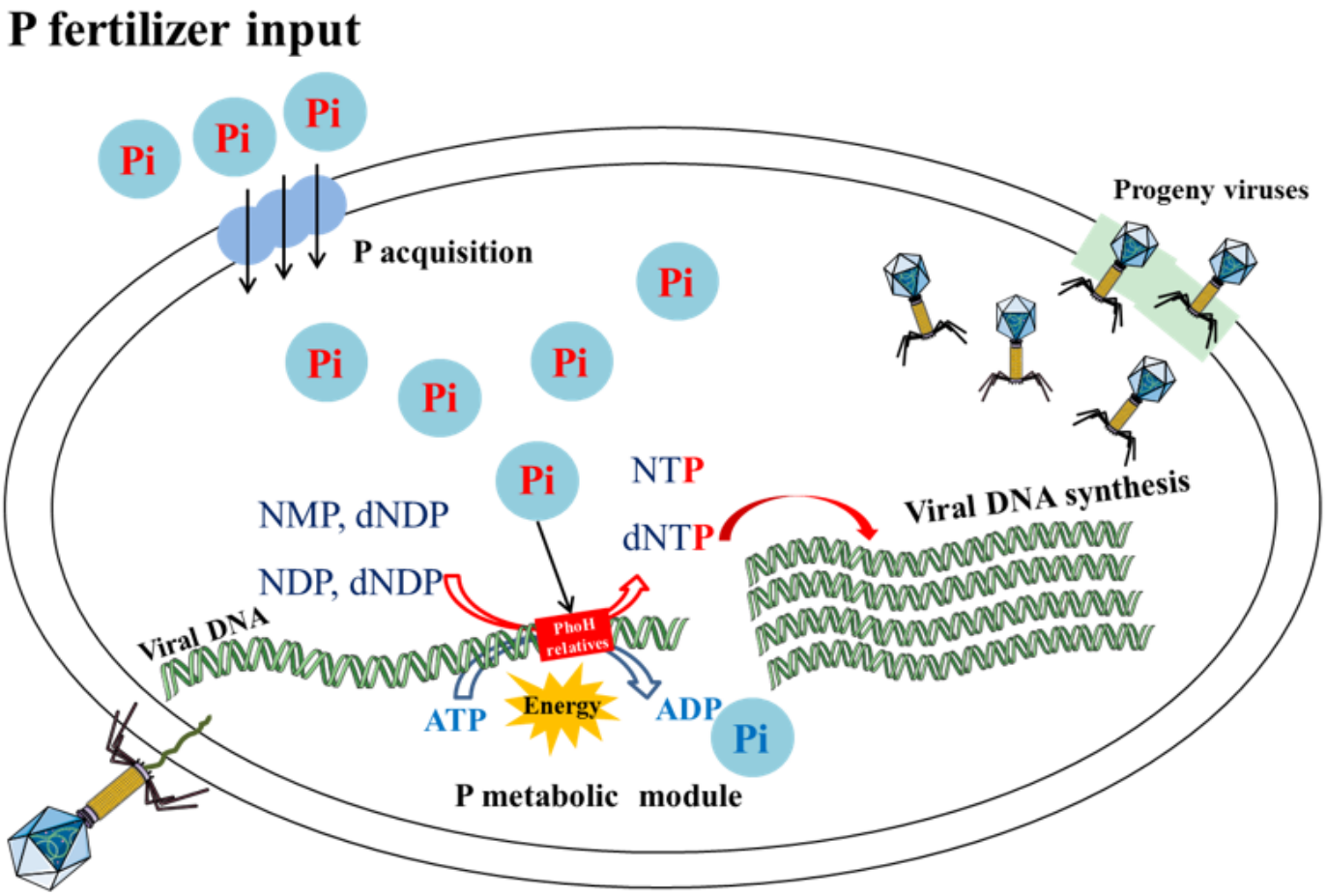

Figure 6

Conceptual model of the P metabolism module of viruses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingInformation.docx