Identification of hidden N4-like viruses and their interactions with hosts in global metagenomes

Kaiyang Zheng
Ocean University of China

Yantao Liang
Ocean University of China

David Paez-Espino
Lawrence Berkeley Laboratory: E O Lawrence Berkeley National Laboratory

Sijun Huang
South China Sea Institute of Oceanology Chinese Academy of Sciences

Xiao Zou
Qingdao Center Hospital: Qingdao Center Medical Group

Chen Gao
Ocean University of China

Yong Jiang
Ocean University of China

Hui He
Ocean University of China

Cui Guo
Ocean University of China

Hongbing Shao
Ocean University of China

Hualong Wang
Ocean University of China

Yeong-Yik Sung
Universiti Malaysia Terengganu

Wen-Jye Mok
Universiti Malaysia Terengganu

Li-Lian Wong
Universiti Malaysia Terengganu

Yuzhong Zhang
Ocean University of China

Jiwei Tian
Ocean University of China

Nianzhi Jiao

Xiamen University

Curtis A. Suttle

The University of British Columbia

Jianfeng He

Polar Research Institute of China

Andrew McMinn

University of Tasmania

Min Wang (✉ mingwang@ouc.edu.cn)

Ocean University of China - Yushan Campus https://orcid.org/0000-0002-2357-589X

Research

Keywords: N4-like viruses, pangenome, virulent factors, viral tRNA, horizontal gene transfer

Posted Date: June 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-539338/v1

To Microbiome

Identification of hidden N4-like viruses and their interactions with hosts in global metagenomes

Kaiyang Zheng[1,2*]; Yantao Liang[1,2*†]; David Paez-Espino[3,4*]; Sijun Huang[5*]; Xiao Zou[6]; Chen Gao[1,2]; Yong Jiang[1,2]; Hui He[1,2]; Cui Guo[1,2]; Hongbing Shao[1,2]; Hualong Wang[1,2]; Yeong Yik Sung[2,7]; Wen Jye Mok[2,7]; Li Lian Wong[2,7]; Yuzhong Zhang[1,8]; Jiwei Tian[9]; Nianzhi Jiao[10]; Curtis A. Suttle[11]; Jianfeng He[12†]; Andrew McMinn[1,13†]; Min Wang[1,2,14†]

1 College of Marine Life Sciences; Institute of Evolution and Marine Biodiversity; Frontiers Science Center for Deep Ocean Multispheres and Earth System; Key Lab of Polar Oceanography and Global Ocean Change, Ocean University of China, Qingdao 266003, China.

2 UMT-OUC Joint Centre for Marine Studies, Qingdao 266003, China.

3 DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

4 Mammoth Biosciences, Inc., South San Francisco, CA, USA

5 CAS Key Laboratory of Tropical Marine Bio-resources and Ecology; South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, Guangdong, China

6 Qingdao Central Hospital, Qingdao 266042, China

7 Institute of Marine Biotechnology, Universiti Malaysia Terengganu (UMT),

21030, Kuala Nerus, Malaysia.

8 State Key Laboratory of Microbial Technology; Marine Biotechnology Research Center, Shandong University, Qingdao 266237, China

9 Key Laboratory of Physical Oceanography; Ministry of Education; Ocean University of China, Qingdao 266100, China

10 Institute of Marine Microbes and Ecospheres; State Key Laboratory of Marine Environmental Sciences, Xiamen University, 361005, China

11 Departments of Earth; Ocean and Atmospheric Sciences; Microbiology and Immunology; Botany and Institute for the Oceans and Fisheries, The University of British Columbia, Vancouver, British Columbia BC V6T 1Z4, Canada

12 SOA Key Laboratory for Polar Science, Polar Research Institute of China, Shanghai 200136, China

13 Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia

14 The Affiliated Hospital of Qingdao University, Qingdao 266003, China

*These authors contributed equally to this work.

†Corresponding authors. Email: liangyantao@ouc.edu.cn (Y. L.); hejianfeng@pric.org.cn (J.H.); andrew.mcminn@utas.edu.au (A.M.); mingwang@ouc.edu.cn (M. W.)

Abstract

Background: N4-like viruses, with specific genomic features and propagation signatures, comprise a unique viral clade within the *Podoviridae* family. N4-like viruses are commonly characterized by the N4-like major capsid protein (MCP) and a giant virion-encapsulated RNA polymerase (N4-like RNAP) with a size of approximately 3,500-aa, which is the largest viral protein so far described. To date, our understanding of N4-like viruses is largely derived from 80 viral isolates that infect bacteria. Thus, it is necessary to expand the diversity of N4-like viruses in culturing-independent methods.

Methods: A Hidden-Markov-Module based method was designed based on two characterized N4-specific marker genes, major capsid protein and N4-like virion-encapsulated RNA polymerase. Viral sub-clades were classified based on the monophyly presented in phylogenic tree and the results of pangenome analysis. Further analysis assessed different distribution patterns, genomic properties, hosts' metabolism reprogramming potentialities, significance of viral tRNA and horizontal gene transfer landscape.

Results: We identified 1,000 N4-like virus sequences from genomes and metagenomes representing diverse habitats from around the world. N4-like viruses have been classified into 27 sub-clades and detected in almost all habitats from pole to pole, including some novel habitats, such as oral mucosa and Antarctica. Virulent factors might be crucial for some human-associated N4-like viruses to reprogram the metabolism of host cells and mediate their pathogenic ability through horizontal gene transfer. From the pangenome analysis, the protein diversity was expended over 7-fold and 17 conserved house-keeping genes were identified. Transcriptional compensation of tRNA indicates that producing progeny virion might be the main significance of viral tRNAs. From the horizontal gene transfer network, some N4-like viral sub-clades were observed that potentially infect some important human pathogens, such as *Campylobacteria* and *Veillonella*, which have not been considered as potential hosts of N4-like virus or even any virus.

Conclusion: This study expands the knowledge of N4-like viruses via global metagenomic datasets, reveals the novel ecological and genomic signatures of these viruses and will provide the backbone for further N4-like virus studies.

Keywords: N4-like viruses, pangenome, virulent factors, viral tRNA, horizontal gene transfer.

Background

Viruses that infect bacteria, known as bacteriophages, could be the most abundant and diverse life entities in the biosphere, with an estimated global population exceeding over 10^{31} [1]. Viruses affect microbial community structure and impact biogeochemical cycles by lysing their host cells and releasing nutrients into the microbial food web (named as “viral shunt”), mediate the virus-host co-evolution through horizontal gene transfer and manipulate the host’s metabolic pathways during infection by expressing viral-encoded auxiliary metabolic genes (vAMGs) [2, 3]. Bacteriophages have been found in a wide range of habitats, including the internal environment of macro-organisms, terrestrial and aquatic area and some extreme environments. They have even been detected in glacier ice, abyssal sediments and fossilized stool specimens from the Middle Ages [4].

Though diverse morphologies are observed in bacteriophages, head-tail caudoviruses still comprise the majority of isolated bacteriophages, according to the public database of 2021 (NCBI virus). The number of species in each family within the Caudovirales is very different. *Siphoviridae* is the largest family, containing almost half of the reported phages, while *Podoviridae* comprises 12 % of reported phages [5]. Currently, four groups of short-tail bacteriophages, T7-like, phi29-like, P22-like and N4-like have been identified [6]. The isolation, genomic features and ecological landscape analysis of SAR11 viruses and SAR116 viruses, suggests that these short-tail bacteriophages might be the most abundant entities in the ocean, or even the entire

Earth [7-9]. Thus, the diversity of short-tail phages might well be underestimated.

N4-like viruses, a lineage of *Podoviridae* with Escherichia virus N4 as the archetype, exhibit uniquely conserved features in their virion structure [10], genome architecture [11] and progression of viral gene expression [11], which is significantly different from T7-like autographiviruses [12]. N4-like viruses are commonly characterized by the N4-like major capsid protein (MCP) and a giant virion-encapsulated RNA polymerase (N4-like RNAP) with a size of approximately 3,500-aa, which is the largest viral protein so far described [13]. In addition, N4-like viruses are the only viral group that transcribes early viral genes without the RNAP of host cells [14, 15]. In the lifecycle of N4, three viral encoded RNAPs (N4-like RNAP, T7-like RNAP1 and RNAP2) play essential roles at different stages of viral propagation [6, 16-21]. This unique transcription program of viral genes highlights that N4-like viruses comprises a special clade in the virosphere, indicating a specific phylogeny in viral evolutionary history.

The first N4-like virus infecting *Escherichia* was isolated in 1967 [12], and 80 N4-like viral genomes have since been reported and published in public databases or the literature. About 54% of these were isolated from *Pseudomonas* (17 isolates), *Escherichia* (13 isolates) and *Roseobacter* (13 isolates). For exclusively marine bacteria, nearly all N4-like viruses were roseophages or vibriophages. Only one N4-like virus infecting *Pseudoalteromonas* has so far been published in GenBank (pYD6-A). Similarly, most host-associated N4-like viruses were found to infect either *Escherichia*, or pathogenic *Pseudomonas*. Isolated N4-like viruses infecting other bacteria were relative rarely reported, suggesting that the investigation of N4-like viruses based on culturing method could be biased and narrow. Thus, the diversity of N4-like viruses has not yet been systematically canvassed. Fortunately, vast numbers of viruses can be identified by bioinformatics analysis of high-throughput sequencing and rapidly growing viral metagenomes. Recently, deep mining of genomes of several essential viral groups (such as giant virus, filamentous phages and ssRNA phages) [22-24], has extended our understanding of global viral diversity and potential linkages among viruses and hosts.

Here, we report on the neglected diversity of 1,000 N4-like viruses (including 920 metagenomic assembled genomes) from GenBank (2021.01) and IMG/VR (2020.10) [25]. These viral genomes were detected and filtered from over 2.4 million viral contigs applying hidden Markov-module [Supporting information Fig. S1]. These genomes reveal that N4-like viruses are far more diverse, widespread, and ecologically ubiquitous than previously appreciated. Our result provided a robust baseline to further mine their potential impacts across multiple hosts and habitats.

Results and discussion

N4-like virus (N4LV) are highly diverse and infect a wide range of bacteria

The metagenomic methods used to detect and screen putative N4LVs to evaluate the global diversity of N4LVs, was based on previous studies [22-24]. This led to the recovery of 1,000 N4LVs, including 80 isolates [Supporting information Tab. S1] and

920 N4LV metagenome-assembled genomes (N4LVMAGs) (see the Supplementary Materials). These genomes were screened out from over 2.4 million viral contigs, belonging to 6,155 metagenomes published in the Integrated Microbial Genomes/Viral Resources (IMG/VR v.3). Using an approach that relied on conserved N4LVOGs, 444 reference and high-quality genomes (labeled as 'high-quality', based on the information provided by IMG/VR) were found, while others were genomic fragments (543) or artificial concatenated sequences (12). The assembly length of viral contigs ranged from 8,393 bp to 198,756 bp with the G+C content ranging from 24.45 to 64.73%. As N4-like MCP and N4-like RNAP are typical marker genes of N4LVs, a phylogenic tree was constructed based on these two genes [Fig.1]. The phylogenic tree was expanded from 80 to 1,000 viral genomes, which can be divided into 27 putative N4LV genera or subgenus-level sub-clades (N4LVSCs), according to their monophyly in phylogeny as well as the presence or absence of N4LVOGs, with most of them being novel. This increases the diversity of the N4LVs by 12.5-fold. Among them, only eight N4LVSCs (N4LVSC2, N4LVSC7, N4LVSC9, N4LVSC11, N4LVSC20, N4LVSC24, N4LVSC25 and N4LVSC27) existed as viral isolates.

For the N4LVSCs containing isolates, the phylogenic distance of some N4LVs is not close, which might reflect the bias caused by different hosts-specificity. For instance, there are eight isolated N4LVs infecting *Vibrio* that all located in the same branch of the tree, which was closer to the root than other branches. Vibrio phage vB_VspP_SBP1 (the only member in N4LVSC2) in particular, has a special location on the tree that indicates it could be much more similar to the common ancestor of N4LVs or underwent a novel phylogenic progress compared to the others. Some isolated N4LVs, including N4LVs infecting *Acinetobacter*, *Pectobacter* and *Pseudoalteromonas,* are only located on the first branch of the phylogenic tree but this could be resulted from the limited number of isolates of these hosts. A similar phenomenon was observed in the N4LVs infecting *Pseudomonas* and *Escherichia* as that of *Vibrio*. Most N4LVs are located in N4LVSC9 (n = 16) and N4LVSC27 (n = 12), except for two that are located in N4LVSC27 (Pseudomonas phage inbricus and ZC08) and one that is located in N4LVSC20 (Escherichia phage Pollock). The N4LVs infecting Rhodobacteraceae (*Roseobacter*, *Dinoroseobacter*, *Roseovarius*, *Sulfitobacter* and *Ruegeria*) are all located in N4LVSC25. In fact, large numbers of isolated N4LVs are located in N4LVSC27 (n = 31), indicating that the diversity of this sub-clade is higher than others. The isolated N4LVs in this sub-clade infects a variety of hosts, including *Achromobacter*, *Delftia*, *Erwinia*, *Escherichia*, *Klebsiella*, *Pseudomonas*, *Shigella*, *Sinorhizobium* and *Xanthomonas*. Many of these hosts are pathogenic. Thus, the positions of different N4LVs on the phylogenic tree could be correlated with the corresponding hosts that they could infect.

There were 19 N4LVSCs without isolates in the species tree, which suggests that the diversity of N4LVs, based on isolation, is likely to be incomplete. The expansion of the number of N4LVs mined from metagenomic datasets is significant. In IMG/VR, the linkages between N4LVMAGs and their putative hosts were predicted by CRISPR, integrated provirus, and genomic similarity with hosts [25-26]. Unfortunately, predicted host information is still lacking and only 75 N4LVMAGs have been predicted successfully. N4LVMAGs infecting Uhrbacteria are included in N4LVSC11. Uhrbacteria is a candidate bacteria phylum belonging to the superphyla Parcubacteria [27], which has not so far been cultured. Fourteen N4LVMAGs infecting Oceanospirillales were detected in N4LVSC13. Most Oceanospirillales taxa have been regarded as only occurring in marine habitats, that have often been enriched by oil contaminated areas and possess the ability to degrade long-chain alkanes [28]. The marine environment could be the primary habitat for those N4LVs in N4LVSC12 N4LVSC13, indicating that these two N4LVSCs could be an undiscovered viral lineage of marine origin. In addition, N4LVMAGs infecting *Thioglobus,* an important marine anaerobic sulfur-oxidizing bacterium [29], dominated in N4LVSC5 and N4LVSC13, from which no virus has yet been isolated. Half of the N4LVs in N4LVSC16 have been identified from human-associated viromes. Nine N4LVMAGs have been predicted to infect *Campylobacter*, an important human pathogenic microbe growing under strictly anaerobic or microaerobic conditions [30]. Over 100 viruses infecting *Campylobacter* have been isolated, but podovirus has not. Similarly, N4LVMAGs predicted to infect *Neisseria* were found in N4LVSC19, a microbe widespread in human/animal-associated environments. *Neisseria* is also an important human pathogen, with some species (eg. *Meningococcus*) being able to infect the mucosal surface of the oropharynx without any symptoms but occasionally triggering invasive meningococcal disease [31-32]. So far, isolated viruses infecting *Neisseria* are still lacking and only one prophage has been reported [33]. Each N4LVSC possesses a signature correlated with their host and habitat, which has been reflected on their phylogenic statuses.

N4LVs are globally distributed and contain diverse viral-encoded auxiliary metabolic genes (vAMGs)

N4LVs are globally widespread and comprised of several sub-clades. In the marine environment, N4LVSC11 and N4LVSC13 seem to be the dominant sub-clades, being observed widely in marine environments including polar areas (Antarctica, Southern Ocean and North Pacific Ocean) [Fig. 2 A]. In addition, N4LVSC25, the sub-clade that includes the majority of marine N4LV isolates such as the N4-like roseophages, are all from near coastal areas. This distribution of marine N4LVs is unexpected, since previous reports were mainly associated with N4-like roseophages. Marine N4LVs comprise nearly half the diversity of N4LVs (45.5%), while others are from host-associated origins (human, animal and plant) (28.6%), waste (11.5%), freshwater (9.8%), saline or alkaline environments (5.9%) and sediment (0.9%). In addition, 55 N4LVMAGs, belonging to nine N4LVSCs, originated from polar environments, including N4LVSC3 (Antarctica: Lake Vida), N4LVSC9 (Antarctica: Lake Vida, Ace Lake, Organic Lake; Greenland; Beaufort Sea), N4LVSC11 (Antarctica: Rauer Islands), N4LVSC13 (Antarctica: Ace Lake), N4LVSC16 (Antarctica: Lake Vida, Ace Lake; Beaufort Sea), N4LVSC18 (Antarctica: Rauer Islands), N4LVSC19 (Antarctica: Ace Lake), N4LVSC25 (Antarctica: Lake Vanda, Rauer Islands), N4LVSC27 (Antarctica: Lake Vanda). This distribution pattern of N4LVs in cold environments has been previously observed [21] but the reason why they are so prevalent there is not clear.

Over two hundred N4LVMAGs associated with humans have been found and these are mainly from viromes associated with the digestive (oral or gut) or respiratory systems [Fig. 2 B]. Digestive systems are a eutrophic environment that contained 148 N4LVMAGs of nine N4LVSCs. The proportion of different N4LVSCs in the mouth and gut is different. The dominant sub-clades in the mouth, such as N4LVSC19 and N4LVSC16, are probably only minor components in the human gut. Some members of these two N4LVSCs have been predicted to infect either *Campylobacter* or *Neisseria*, both of which could colonize oral mucosa. Eight different N4LVSCs were derived from the viromes of the human gut [Fig. 2B], suggesting a high diversity of N4LVs in this eutrophic and hypoxic environment. N4LVSC7 and N4LVSC23 are the dominant sub-clades, contributing 56% of the total N4LVs in the human gut. Host information of these N4LVSCs is lacking, indicating that more complex microbes inside the human body might be infected by unknown N4LVs. The diversity of N4LVs in the human respiratory system was much lower than that of the digestive system. Only four N4LVs, classified into two N4LVSCs, have been identified from viromes from the lungs. It is possible that widespread N4LVs might have a mutualistic or coevolutionary relationship with complex microbial communities, mediated by vAMGs located in their genomes.

Viruses are able to reprogram metabolic pathways by expressing vAMGs that influence the physiological behavior of the host and further mould microbial community structure [34]. A total of 627 genes have been confirmed as vAMGs that can be classified into 13 types [Fig. 2 C]. Notably, virulence factors are the most frequently observed vAMGs of N4LVs. Virulence factors are most abundant in N4LVSC19, followed by N4LVSC16. Most virulence factor-related genes tend to be carried by host-associated N4LVs. Eleven virulence factors were identified, including *virC1* protein [35], *yopX* protein [36], *yadA* protein [37], GDP-mannose dehydrogenase [38-40], F plasmid-carried bacterial toxin [41], *virE* protein [42], *ompA* protein [43], *ydaS* antitoxin protein [44], *zeta* toxin protein [45], *sdr* proteins [46] and bacterial neuraminidase [47]. The *yadA* protein is the most virulent factor (n = 112) that is encoded by 40 N4LVs (including three isolates: Salmonella phage FSL_SP-058, FSL_SP-076 and Erwinia phage vB_EamP-S6), which mainly exist in N4LVSC19 (n = 30). Bacteria use *yadA* protein to infect their host cells via a process of cell adhesion, making the bacteria harmful and infectious to the host organisms [37]. The *yopX* protein was found in 28 N4LVMAGs of N4LVSC16. This protein was first discovered in outer proteins produced by *Yersinia*, which is exported by the type III secretion system upon bacterial infection of host cells. The type III secretion system is encoded on a virulence plasmid and is necessary for the survival and replication of the bacterium within the host's lymphoid tissue [36] and might promote the exacerbation of glandular plague if *Yersinia pestis* is impacted by infection with these kinds of N4LVs. These two proteins are the most abundant virulent factors in N4LVs (78.9%), indicating the crucial role they play in reprogramming the metabolism of their hosts, especially for parasitic bacteria. Three vAMGs are referred to the electron transfer and cellular respiration, including the Fe-S cluster biosynthesis protein (4), thioredoxin (32) and ferredoxins (2). They also dominated in host-associated N4LVs (68), excepting N4LVSC16 and N4LVSC19. A significant number of N4LVs had these vAMGs, which belong to N4LVSC23 (29) and N4LVSC25 (17). Given that most of the N4LVs belonging to N4LVSC23 and N4LVSC25 lack known virulent-associated proteins and that

N4LVSC16 and N4LVSC19 lack unknown electron transfer-associated proteins, different strategies of host metabolism reprograming might be applied in different host-associated N4LVs. Our data shows that marine N4LVs might tend to carry less vAMGs than host-associated N4LVs but some types of vAMGs are specific to marine N4LVs. Fat and sulfur metabolism-associated protein regulation factors are dominated by some marine N4LVs, while cellular secretion and cellular signaling related proteins might only occur in marine N4LVs. Marine N4LVs might need to reprogram their bacterioplankton host's metabolism more accurately to maximize their efficiency in resource-poor oligotrophic ocean gyres.

Pangenome of N4LVSCs demonstrated a special viral conserved genetic strategy

Genetic conservation was common in N4LVs. Some N4LVOGs are shared in almost all N4LVSCs while some sub-clades were characterized by some unique N4LVOGs [Supporting information Fig. S2]. We investigated the pangenome of N4LVSCs through relationship between N4LVOG accumulation viral genome accumulation [48]. Similar work has focused on N4-like roseophages [20]. We extended this analysis to all the N4LVs generated in this study. The pangenome analysis was conducted separately on 80 isolated N4LVs [Supporting information Fig. S3] and all 1,000 N4LVs, respectively [Fig. 3 A], producing 642 and 4,951 N4LVOGs and singletons by all-against-all BLASTp. The N4LVOGs increased in abundance 7.7-fold, while species increased about 12.5-fold. This result demonstrates the astringency of N4LVOGs expansion when the number of N4LV genomes was increased. The number of N4LVOGs barely increased after the addition of approximately 700 genomes of

N4LV, illustrating that the genetic diversity of N4LV genes may have reached saturation.

N4LVSCs containing at least 10 members were included to analysis core-genome signature, thus 12 N4LVSCs were included. There are 3,845 N4LVOGs and singletons in N4LVSC, with a large proportion being singletons (3,413). Different N4LVOGs are shared between N4LVSCs with at least 24 being shared (N4LVSC7 and N4LVSC13) [Fig. 3 B]. These shared N4LVOGs contained both core genes and dispensable genes; core genes are shared in all or nearly all N4LVs while the latter are shared in at least two N4LVs. Sixteen N4LVOGs are found in all N4LVSCs [Fig. 3 D], including seven metabolism genes (homologous genes in N4: gp16, gp24, gp39, gp42, gp43, gp44 and gp45), three structure genes (gp56, gp57 and gp59), one packaging gene (gp58) and five conserved genes with unknown functions (gp52, gp54, gp55, gp58 and gp69). These genes in the viral genomes are modular and are located in specific areas of the sense or anti-sense strand. The relative position of each module remains conserved, although it can occasionally be reversed. The terminase large subunit (gp68) is followed by a conserved hypothetical protein (gp69). Three non-modular genes are dispersed in different loci, including the RNAP2 (gp16), AAA-domain ATPase (gp24) and a conserved hypothetical protein (gp52). Unlike the N4-like vRNAP, T7-like RNAP can mediate the transcription of viral middle genes [14]. This might be the reason that the locus of RNAP2 is away from other modular core genes. This gene is followed by the AAA-domain ATPase, a protein participating in the regulation of cellular proteolysis, which synergistically acts with proteasomes and proteasome-like proteases in protein degradation [49-50]. These two genes and DNA replication modules are typical viral early genes of N4LVs that are responsible for nucleotide replication and host metabolism reprogramming. The structure and packaging gene modules are a viral late gene that are responsible for virion mutations. Thus, the conservative genetic arrangement of N4LVs suggests that a similar reproduction strategy could be applied to all N4LVs. Portal protein, tape-measure protein, MCP, and four conserved hypothetical proteins (gp52, gp54, gp55 and gp58) are clustered on a relatively narrow area of the genomes, while a similar situation occurred in the ssDNA binding protein, AAA-domain ATPase, DNA primase, PD-(D/E)XK nuclease and DNA polymerase. Two gene modules are located on a different strand next to the N4-like RNAP.

Core-genome dilution by accumulation of genomes of each N4LVSC provides a general view of the conserved gene group at the sub-clade scale [Fig. 3 C]. Core genes of a sub-clade are shared throughout nearly all members of the sub-clade. N4LVOGs encoded by over 90% viral genomes were regarded here as core genes. The number of identified core genes under the applied threshold in each N4LVSC differed substantially, ranging from four (N4LVSC16) to 52 (N4LVSC20), but that of most N4LVSCs are between 10 and 30. A high percentage of core genes in N4LVSC20 may imply its highly conserved phylogenic state. N4LVSC16, by contrast, has the lowest percentage of core genes, that includes only MCP (gp56), tape-measure protein (gp57) and two conserved hypothetical proteins (gp58 and gp69). This sub-clade included 45 N4LVMAGs, although no N4LVs have so far been isolated. This result is unusual because N4LVSC16 has a monophyly in the phylogenic tree [Fig. 1]. The overlap of N4LVOGs of N4LVSC16 among N4LVSCs, and sub-clade-specific N4LVOGs, also indicates the special features of this sub-clade that might possess more complex genetic types than other N4LVSCs [Fig. 3 B and Supporting information Fig. S2].

The pangenome has a large genetic pool driving the evolutionary processes of N4LVs, since the capacity to increase diversity of a viral clade is limited by its gene pool [51]. The core-genome of N4LVs are the house-keeping genes of this viral clade that is conserved in both function and locus, forming a crucial backbone of N4LVs. These house-keeping genes are present in N4LVs infecting different hosts, suggesting all N4LVs might have evolved from a viral common ancestor.

Compensation of gene transcription by N4LV carried tRNA

The tRNA carried by viruses promotes translation efficiency of viral genes, as the tRNA pool of viruses compensates the tRNA pool of their hosts during the viral propagation. Synergy of codon usage (CU) between a virus and corresponding host could be a major force mediating co-evolution of virus-host [52]. However, the viral tRNA genes and CU difference between them is more sophisticated, as similar CU patterns could be negative for the propagation of virions in a cell [53]. The enhancement of viral tRNA to each gene in viral genomes is not even, since the translation of certain genes can be enhanced with a bias by the viral tRNA pool [54-55]. For the 444 high-quality N4LVs, more than half of the genomes (258) contained tRNA genes. The average and highest number of tRNA genes carried by N4LVs was 3 and 23, respectively (see the Supplementary Materials). The N4LVMAGs containing tRNA mostly originated from host-associated sub-clades (133), followed by those of marine (59) and freshwater (21) origin. Apparent bias among N4LVSCs was also observed, members of N4LVSC19 occupied the largest portion of tRNA-carrying N4LVs (44), followed by N4LVSC27 (39) and N4LVSC23 (35). Host-associated N4LVs likely carried more tRNA to enhance transcription of viral genes, which is particularly notable in N4LVSC19.

A similar protocol was used t to investigate the pattern of viral tRNA compensation in N4LVs. The biased tRCI (tRNA compensation index) (considering only the viral tRNA pool) was used to characterize tRNA compensation of viral gene translations. The calculated tRCI ranged from 23.9 to 0.01; higher values indicating higher enhancement of the corresponding viral gene. For the isolated N4LVs, only the first bin passed the hypergeometric distribution test ($p < 0.05$), indicating enrichment of this bin was effective. The proteins with known function in the first bin were classified according to their functions [Supporting information Fig. S4]. Four groups were assigned that included replication related proteins, structure/packaging related proteins, lysis related proteins and vAMGs. A similar result was observed between tRCI and bias tRCI. Replication related proteins contributes more than half of the proteins with known functions in both calculations (63.4% and 63.8%, respectively), which is followed by structure/packaging related proteins (28.4% and 22%, respectively) and lysis related proteins (6.7% and 9.4%, respectively). Only a small number of vAMGs benefit from viral tRNA (1.5% and 4.7%, respectively).

The analysis of bias tRCI was extended to all 258 tRNA-carrying N4LVs. Proteins with known functions were chosen as subjects to assess their enrichment efficiency. Three bins (Bin 2, Bin 3 and Bin 9) passed the hypergeometric distribution test (p <

0.05) [Supporting information Fig. S5], indicating that the subject enrichment of these bins is significant. Given that the hypergeometric distribution test of Bin 1 failed and there is a relatively low tRCI result for Bin 9, only Bin 2 and Bin 3 were used for further analysis. The proteins with known functions in these two bins were classified into 10 types, including transcription regulation related proteins, nucleic acid synthesis, peptides modification related proteins, DNA synthesis related proteins, packaging related proteins, viral absorption related proteins, structure related proteins, vAMGs, lysis related proteins and others (conserved viral protein with unknown function) [Fig. 4]. Generally, the proteins of nucleic acid synthesis comprise the largest proportion of these viral tRNA-enhancement proteins (23%). This is followed by transcription regulation related proteins (17%) and DNA synthesis related proteins (15%). Thus, the viral replication-associated genes benefit the most from viral tRNA.

The second viral tRNA-enhancement protein group varied by sub-clade. The proportion of lysis-related proteins in N4LVSC7 and N4LVSC20 was 15% and 18% respectively, suggesting transcription of lysis-related proteins might be enhanced by the viral tRNA of these N4LVs in the late period of viral propagation. Peptide modification-related proteins comprised a large proportion of the proteins in N4LVSC23 (15%). The radical SAM protein also comprise a large proportion (44%) in this protein group, followed by serine/threonine protein phosphatase (19%) and ATP-dependent zinc metalloprotease (19%). Diverse reactions are catalyzed by radical SAM proteins, including unusual methylations, isomerisation, sulphur insertion, ring formation, anaerobic oxidation and protein radical formation [56]; so they play a vital role in viral propagation after injection of virions.

Viral tRNA provided only minor beneficial to vAMGs according to our data, with only 2% of vAMGs being enhanced by viral tRNA. Hence, the manipulation of the host's metabolism might not be the main purpose of the tRNA carried by N4LVs. Given that almost the entire viral tRNA pool is used to enhance the replication of virions, improvement of the proliferation ratio to generate more progeny virion could be of significance for this viral tRNA, although this result is impacted by currently available annotation information.

Linkages between N4LVs and microbial hosts through horizontal gene transfer (HGT)

Horizontal gene transfer (HGT) is prevalent in microbial communities. The patterns of gene exchange among various groups of bacteria have been described [57-59]. A total of 5,856 genes of 907 N4LVs have been identified as HGT candidates (see the Supplementary Materials). HGT network illustrates the linkages between N4LVs and their putative hosts, which included 85 families under 42 bacteria orders [Fig. 5]. Proteobacteria seemed to be the most common host of N4LVs, comprising over 97% of total linkages. Alphaproteobacteria is the largest host group (42%), followed by Gammaproteobacteria (35%), Epsilonproteobacteria (15%) and Betaproteobacteria (5%). The HGT linkages associated with Firmicutes and Bacteroidetes comprise less than 2% and 1%, respectively. The putative hosts from the ten orders contribute over 95% of the linkages, including Rhizobiales (37.1%), Campylobacterales (14.8%), Oceanospirillales (11%), Enterobacterales (8.7%), Pseudomonadales (8.3%), Alteromonadales (5.2%), Burkholderiales (3.9%), Rhodobacterales (3.4%), Veillonellales (1.7%) and Chromatiales (1.1%). As no N4LVs infecting rhizobium have yet been isolated, the frequent genetic exchanges among N4LVs and Rhizobiales are unexpected. Rhizobiales are crucial microbes in the rhizosphere, which plays a vital role in nitrogen fixation [59]. Eleven families within Rhizobiales are present in the network, including Phyllobacteriaceae, Brucellaceae, Rhizobiaceae, Hyphomicrobiaceae, Aurantimonadaceae, Chelatococcaceae, Salinarimonadaceae, Methylocystaceae, Beijerinckiaceae, Xanthobacteraceae and Bradyrhizobiaceae. Over half of HGT candidates within this order are from Phyllobacteriaceae (1179), followed by Hyphomicrobiaceae (482) and Brucellaceae (432). The high number of HGT linkages between N4LV and rhizobia might suggest a high diversity of N4LVs in the rhizosphere and phyllosphere. The N4LVs infecting rhizobia contain a large number of host-originated genes, indicating a special co-evolutionary process through genetic exchange. Plant-associated rhizobia from saprophytes and endosymbionts occupy an essential ecological niche [61-62]. The interactions between N4LVs, rhizobia and plants are complex and not well known. By contrast, Rhodobacterales, another common putative host order within Alphaproteobacteria, that includes most marine-associated N4LVs that infect Alphaproteobacteria (roseophages), comprise only 3.4% of all HGT linkages. Marine-originated N4LVs are nearly all roseophages but the proportion of HGT linkages referred to these N4-like roseophages is small. Approximately 14% of HGT linkages are associated with Campylobacterales (865). The N4LVMAGs grouped with Campylobacterales has hardly any HGT linkages with others. Oceanospirillales are probably the commonest host lineage of N4LVs in the marine environment, comprising 11% (645) of total HGT linkages, including the families in this order (Halomonadacae, Oceanospirillacae, Alcanivoracaceae and Kangiellaceae). Most Oceanospirillales-associated HGT linkages are found in Halomonadacae (n = 324) and Oceanospirillacae (n = 316).

Forty-two putative bacteria orders were linked by this N4LV-mediated network and diverse HGT patterns, corresponding to different N4LVSCs, were observed. For instance, the HGT linkages of Campylobacterales, Veillonellales and Uhrbacteria form three different groups in the network, which are from N4LVSC16, N4LVSC19 and N4LVSC 11, respectively. Most HGT linkages were observed in N4LVSC7 (10.9%), which contains 39 N4LVs, including five isolated N4LVs (Acinetobacter phage Presley and vB_ApiP_XC38; Pectobacterium phage vB_PatP_CB1, vB_PatP_CB3 and vB_PatP_CB4). This is followed by N4LVSC16 (9.8%) and N4LVSC13 (9.3%). Most N4LVSC16-associated HGT linkages are linked to *Campylobacter*. Similarly, N4LVSC19 is closely related to Veillonellales and Neisseriales. Nearly all N4LVSC27-associated linkages are related to Rhizobiales and Enterobacterales. The two marine-derived N4LVSCs (N4LVSC12 and N4LVSC13) are closely related to Oceanospirillales. Many of the N4LVSC13-associated HGT linkages are also from Rhizobiales. N4LVSC25 and N4LVSC26 mediated linkages between Rhizobiales and Rhodobacterales, which contained all isolated N4-like roseophages. Similar N4LV-mediating linkages were observed between other bacteria, such as Enterobacterales and Neisseriales (linked by N4LVSC16), Enterobacterales, Pseudomonadales and

Alteromonadales (linked by N4LVSC25), Oceanospirillales, Uhrbacteria (linked by N4LVSC11), Burkholderiales and Neisseriales (linked byN4LVSC19).

Many of the important pathogenic bacteria were detected in the HGT network, which are all from Enterobacterales and Vibrionales. Nine N4LVs were associated with *Vibrio*, including three isolates (*Vibrio* phage phi1, JA-1 and VCO139) and six N4LMAGs. A number of N4LVs are associated with Enterobacterales, mostly from Enterobacteriaceae (118), Erwiniaceae (310), and Yersiniaceae (50). Seventeen isolated N4LVs are linked with Enterobacteriaceae in the network, most of which infected *Escherichia* (Escherichia phage vB_EcoP_Bp4, EC1-UPM, vB_EcoP_PhAPEC7, ECBP1, PMBT57, St11Ph5, PhAPEC5, vB_EcoP_G7C and N4), *Salmonella* (Salmonella phage FSL_SP-058 and FSL_SP-076) and *Pectobacterium* (Pectobacterium phage vB_PatP_CB1, vB_PatP_CB3 and vB_PatP_CB4). A total of 44 N4LVs are associated with Yersiniaceae through HGT, indicating that N4LVs infecting *Yersinia* might be diverse and abundant, even though no isolated N4LV infecting this lethal pathogen has yet been reported.

The complex genetic exchange between N4LVs and their putative hosts was reflected in this HGT network, demonstrating the genetic exchange landscape between N4LV and microbes. However, HGT has rarely been observed in Pelagibacterales (only one linkage of Pelagibacteria) and Cyanobacteria (only one for *Synechococcus*), suggesting that N4LVs infecting these two abundant marine microbes might be rare. Thus, while the habitats that N4LVs occupy were diverse, the oligotrophic pelagic zone might not be a natural habitat for them. Some eutrophic areas, such as animal related internal environments, soil, sewage and coasts might be their preferred habitats. However, polar area is also a potential habitat of N4LVs based on our result and the reason is still unclear. Diverse proteobacteria seem to be the only hosts for these N4LVs, indicating that they might have evolved and differentiated with proteobacteria during their evolutionary process. Given that isolated N4LVs occupy only a small proportion of the HGT network, this result sheds light on the diverse N4LV-mediated genetic exchange patterns within microbes.

Conclusion and remarks

This investigation, drawing on many previous reports, has expanded the known diversity of N4LVs. It provides a solid foundation and resource for further investigation of the different effects that N4LVs have on microbial ecosystems. It will also deepen our understanding of the evolution and biogeochemical role of these viral clades. N4LVs are a series of conserved viruses with stable genomic architecture, which might be essential for their niche in the virosphere. The general evolutionary concept is not appropriate for viruses, as common ancestor theory is not applicable to these life entities. It seems that viruses evolved with cellular organisms at the beginning of life and new viruses were continuous being generated with the differentiation of cells, tissues and organisms. The origin of viruses spans across billions of years after life first appeared on earth. Viruses commonly possess high variation rate, while some house-keeping genes are remarkably stable in different viral clades. This makes it possible to unveil the cryptic viral phylogenic evolution through investigation of these viral house-keeping genes. Furthermore, the genetic linkages among viruses, prokaryote and eukaryotes might provide a hidden the key to understand the diverse patterns of life now.

Given that no N4LV infecting Cyanobacteria and Pelagebacteria have yet been reported and hardly any linkages in the HGT network were observed, the reasons for the restricted host spectrum of N4LV remains unclear. As many isolated or metagenomic assembled N4LVs are from eutrophic environments, this suggests that greater resources might be required for the propagation of N4LVs. However, N4LVs are also distributed in polar area, which are often oligotrophic, extreme environments. Being able to colonise different extreme environments indicates the strong potential for environmental adaptation of N4LVs, which reflects their successful survival strategies. It is understood that N4LVs are pervasive in the internal human environment, indicating that N4LVs are likely to be an as yet hidden but crucial component of the human body ecosystem. As deadly viruses have been rampant in human society for thousands of years, these human-associated viruses carrying virulent factors deserve to be investigated to a much greater degree. The deeper we expand our knowledge of viruses, the closer we get to the essence of life.

Methods

Generate module to detect N4LVs in metagenomics

Marker genes were determined based on the principle that a particular gene should be relatively conserved and present as a single copy in all putative N4LVs. Hence, the N4-like major capsid protein (MCP) and N4-like virion-encapsulated RNA polymerase (N4-like RNAP) were recognized as significant characteristics of N4LVs [14-15]. As these two genes are concentrated on fragments of approximately 20 kbp, searching and filtering based on these two genes should reveal as many N4LVs as possible. A similar protocol to the previous one was used here to find putative N4LVs in pre-assembled IMG/VR datasets. [22-24, 63]. Briefly, 80 published N4-like phages in GenBank were selected as the reference sets. The reference N4-like MCPs were aligned by MAFFT (v.7.453) [64] under the global aligning mode. The multiple alignments were used to build initial N4-like hidden-Markov-modules (HMMs) by HMMER3 [65], which were then used to conduct the initial search in the IMG/VR protein datasets [26]. The hmmsearch (E-value < 1e-5) first produced 7279 putative N4-like MCPs. Then potential duplications (over 99% similarity in over 90% region of each alignment) were removed by CD-hit [66], resulting in 7260 representative MCPs. Secondly N4-like HMMs were generated from the results generated in the last step. A second hmmsearch (E-value < 1e-5) produced 27432 putative N4-like MCP.

The N4-like RNAP was searched by the same method described above, resulting in 2546 putative N4-like RNAP (E-value < 1e-5). The N4-like RNAP HMMs were then used to filter all viral contigs containing N4-like MCP. Any contigs excluding N4-like RNAP, and contigs containing more than one copy of two marker genes were removed; this resulted in 920 viral contigs [Supporting information SI_Figure1].

Protein functional annotation and tRNA detection of N4LVs

All N4LV proteins were annotated against non-redundant protein sequences (NR) (2021.01) by Diamond BLASTp (v.0.9.21) (E-value < 1e-5) [66] and Pfam-A (v.33.3) [68] by pfam_scan.pl (v.1.6). The tRNA sequences were detected for all high-quality viral genomes by tRNAscan-SE (v.2.0) [69] under bacteria source and default search mode. The vAMGs were identified based on annotations from Pfam-A.

Construction of species tree

Two marker genes were used to construct the phylogenic tree. Two marker genes of all 1000 N4LVs were aligned by MAFFT (v.7.453) [64] in global aligning mode and trimmed by trimAL (v.1.4) [70] to remove 90% gap region for each alignment). Two proteins of the marker genes were concatenated by SeqKit (v.0.13.2) [71] and transferred to IQ-Tree2 [72] to calculate the maximum-likelihood phylogenic tree under ultrafast mode with the suggested protein module LG+F+R10. The tree was visualized by iTOL [73]. Twenty-seven sub-clades of N4LV (N4LVSC) were assigned based on their monophyly in the species tree and the presence or absence of N4LVOGs.

N4LVOGs detection and Pan-/core-genome analysis

In order to cluster the proteins into different families, all-against-all BLASTp (E-value < 1e-5, query cover > 50%, percentage of identity > 30%) were conducted for all 61769 N4LVOGs of the 1000 N4LVs. Othorfinder2 [74] was used to cluster the N4LVOGs from the results of the all-against-all BLASTp, resulting in 4951 N4LVOG and singletons. All N4LVOGs were mapped to corresponding genomes plotting the pangenome accumulation curve by R. Each additional genome was randomly sampled 100 times and the medians of the boxes (boxes was not displayed for clarity) were linked by curve. In the core-genome analysis, only high-quality genomes of N4LV were included to avoid the bias of using incomplete genomic fragments. Core-gene accumulated curves were plotted by a similar method to the pangenome analysis. Only N4LVSCs containing at least 10 high quality-genomes so that 12 N4LVSCs were included in this analysis. Clade-specific N4LVOGs (unique genes) of the 12 N4LVSCs were visualized with heatmap using R. Twenty-one high-quality genomes of N4LV were selected for the mapping synteny plot of the comparative genomics using ViPTree (v.1.9) [75].

Assessment of translation compensation for N4LV-encoded tRNA

In order to investigate the compensation of gene translation efficiency that is enhanced by viral-encoded tRNA, tRCI (tRNA compensation index) was used to assess such compensation [54]. Host information on the N4LVs is mostly lacking and only available for 74 of the N4LVMAGs. As detailed host information is unknown, a modified method was applied here.

First, tRCI of the isolated N4LVs containing tRNA was calculated in two ways, on whether viral gene translation by tRNA-pool-enhancement significantly depends on the host or not. Results from both methods were compared to determine discrepancy. A CU table of each coding sequence of the related 44 tRNA-carrying N4LVs and their corresponding host genomes, was calculated by EMBOSS. Codon frequency (per 1000 codons) of each tRNA in the corresponding N4LV were retrieved from the CU table of each viral sequence and host genome, separately. For a single viral coding sequence, the continued product (item was ignored if value of corresponding frequency was 0) was calculated through the retrieved codon frequencies in the last step, then divided by the continued product of that of host to produce the $tRCI_1$ of this coding sequence.

The continued product of the corresponding codon frequency of the hosts was then regarded as a constant ($n = 1$ here). This aimed to remove influence of drivers of the tRNA pool in different hosts, then $tRCI_2$ was calculated for the same coding sequence. The two arrays of tRCI produced in last step were normalized by z-scoring and the Pearson's correlation coefficient (R-value) was calculated to characterize the discrepancy between the arrays. A high similarity and correlation were observed (R-value = 0.988), indicating the latter method was reliable.

The 257 tRNA-containing N4LVs (213 high-quality N4LVMGs and the 44 isolated N4LVs) were used to calculated tRCI, based only on their tRNA pool. A total of 20196 biased tRCI of the N4LVs coding sequences were calculated, sorting by their values. These sorted tRCI were equally divided into ten bins ($n = 2020$, approximately). A hypergeometric distribution test was used to assess the enrichment of these bins (enrichment was regarded as effective for a bin if $p < 0.05$), based on proteins with known functions that were subjected to BLASTp against the NR (2021.01). According to the bin of tRCI, the studied gene groups (proteins with known functions) were considered as having an efficient translation enhancement by viral tRNA if the related bins passed the hypergeometric distribution test.

HGT network among N4LVs and putative host

A similar protocol to that previous applied was used to construct the HGT network [22]. Briefly, all proteins belonging to the high-quality N4LVs were subjected to BLASTp against the viral NR datasets (taxonomy 10239) and bacteria NR database (taxonomy 2) (E-value<1e-50, query/subject cover > 50%, percentage of identity > 50%). Best hits within the same N4LVSCs (viral NR database) and orders (Prokaryotic NR database) were removed. Hits with lower E-value in the Prokaryotic NR database compared to those in viral NR database were considered as a candidate HGT. The taxonomic categories of the bacteria hosts with HGT linkage were retrieved from NCBI by TaxonKit [76]. Results with at least 'Order' taxon information were retained to build the network, resulting in 5856 HGT candidates. Twenty-seven N4LVSCs (444 high-quality N4LVs) were used to construct the bacteria-viral HGT linkage network through Gephi (Force Atlas, edge weight 2) [77].

List of abbreviations

Full names	Abbreviations
Basic Local Alignment Search Tool: Protein	BLASTp
Codon Usage	CU
Expectation coefficient	E-value
Horizontal gene transfer	HGT
Integrated Metagenome and Genome/Viruses Resource	IMG/VR
N4-like virus	N4LV

N4-like viral metagenomics assembled genome	N4LVMAG
N4-like viral orthologous group	N4LVOG
N4-like viral sub-clade	N4LVSC
Non-Redundant protein database	NR
Significance coefficient	P-value
Pearson correlation coefficient	R-value
tRNA compensated index	tRCI
Viral Auxiliary Metabolism Gene	vAMG

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The N4-like viral contigs generated in this study are available in extending data 4-5.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (No.2018SDKJ0406-6), National Key Research and Development Program of China (2018YFC1406704 and 201812002), the Fundamental Research Funds for the Central Universities (201812002), Natural Science Foundation of China (No. 41976117 and 41606153), the Key R&D project in Shandong Province (2019GHY112079), and the MEL Visiting Fellowship Program (MELRS1511).

Author contributions:

K.Z., Y.L., J.H., A.M. and M.W. designed research; K.Z., Y.L., D.P.-E., and S.H. performed research; K.Z., C.Gao, Y.J. and H.H. contributed analysis tools and computing scripts; K.Z., X.Z., C.Guo, H.S. and H.W analyzed data; K.Z., Y.Y.S., W.J.M. and L.L.W. visualized data; K.Z., Y.L., J.H., A.M. and M.W. wrote this manuscript; Y.Z., J.T., N.J., D.P.-E., and C.A.S. refined the manuscript.

Reference

[1] Suttle CA. Viruses in the sea. Nature. 2005; 437: 356–361. https://doi.org/10.1038/nature04160

[2] Robinson C, Ramaiah N. “Microbial heterotrophic metabolic rates constrain the microbial carbon pump,” in Microbial Carbon Pump in the Ocean, eds F. Azam, N. Jiao, and S. Sanders (Washington, DC: Science/AAAS), 2011: 52–53.

[3] Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3: 754–766. https://doi.org/10.1038/s41564-018-0166-y

[4] Appelt S, Fancello L, Bailly ML, Raoult D, Drancourt M, Desnues C. Viruses in a 14th- century coprolite. J Appl Environ Microbiol. 2014; 80: 2648–2655. https://doi.org/10.1128/AEM.03242-13

[5] Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020; 18: 125–138. https://doi.org/10.1038/s41579-019-0311-5

[6] Zhao Y, Wang K, Jiao N, Chen F. Genome sequences of two novel phages infecting marine roseobacters. Environ Microbiol. 2009; 11(8): 2055–2064. https://doi.org/10.1111/j.1462-2920.2009.01927.x

[7] Zhao Y, Temperton B, Thrash JC, Schwalbach MIS, Vergin KL, Landry ZC, et al. Abundant SAR11 viruses in the ocean. Nature. 2013; 494:357–360. https://doi.org/10.1038/nature11921

[8] Y. Zhao, Qin F, Zhang R, Giovannoni SJ, Zhang Z, Sun Z, et al. Pelagiphages in

the Podoviridae family integrate into host genomes. Environ Microbiol. 2018; 21(6): 1989–2001. https://doi.org/10.1111/1462-2920.14487

[9] Zhang Z, Qin F, Chen F, Chu X, Luo H, Zhang R, et al, Culturing novel and abundant pelagiphages in the ocean. Environ Microbiol. 2020; 00(00): 00–00. https://doi.org/10.1111/1462-2920.15272

[10] Liu X, Zhang Q, Murata K, Baker ML, Sullivan MB, Fu C, et al. Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. Nat Struct. Mol. Biol. 2010; 17: 830–836. https://doi.org/10.1038/nsmb.1823

[11] Chen F, Lu J. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. J Appl Environ Microbiol. 68: 2589–2594 (2002). https://doi.org/10.1128/aem.68.5.2589-2594.2002

[12] Schito GC, Molina AM, Pesce A, Lysis and lysis inhibition with N4. Giorn Microbiol. 1967; 15: 229–244.

[13] Ma Y, Li E, Qi Z, Li H, Wei X, Lin W, et al. Isolation and molecular characterisation of Achromobacter phage phiAxp-3, an N4-like bacteriophage. Sci Rep. 2016; 6(1):24776. https://doi.org/10.1038/srep24776

[14] Falco SC, Laan KV, Rothman-Denes LB. Virion-associated RNA polymerase required for bacteriophage N4 development. Proc Natl Acad Sci U S A. 1977; 74: 520–523. https://doi.org/10.1073/pnas.74.2.520

[15] Falco SC, Zehring W, Rothman-Denes LB, DNA-dependent RNA polymerase from bacteriophage N4 virions: Purification and characterization. J Biol Chem. 1980;

255: 4339–4347. https://doi.org/10.1016/S0021-9258(19)85670-3

[16] Davydova EK, Santangelo TJ, Rothman-Denes LB, Bacteriophage N4 virion RNA polymerase interaction with its promoter DNA hairpin. Proc Natl Acad Sci U S A. 2007; 104:7033–7038. https://doi.org/10.1073/pnas.0610627104

[17] Willis SH, Kazmierczak KM, Carter RH, Rothman-Denes LB. N4 RNA Polymerase II, a heterodimeric RNA polymerase with homology to the single-subunit family of RNA polymerases. J Bacteriol. 2002; 184: 4952–4961. https://doi.org/10.1128/JB.184.18.4952-4961.2002

[18] Choi M, Miller A, Cho NY, Rothman-Denes LB. Identification, cloning, and characterization of the bacteriophage N4 gene encoding the single-stranded DNA-binding protein. A protein required for phage replication, recombination, and late transcription. J Biol Chem. 1995; 270: 22541–22547. https://doi.org/10.1074/jbc.270.38.22541

[19] Chan JZM, Millard AD, Mann NH, Schafer H. Comparative genomics defines the core genome of the growing N4-like phage genus and identifies N4-like Roseophage specific genes. Front Microbiol. 2014; 5:506. https://doi.org/10.3389/fmicb.2014.00506

[20] Zhan Y, Chen F. Bacteriophages That Infect Marine Roseobacters: Genomics and Ecology. Environ Microbiol. 2018; 21(6): 1885–1895. https://doi.org/10.1111/1462-2920.14504

[21] Zhan Y, Buchan A, Chen F. Novel N4 Bacteriophages Prevail in the Cold Biosphere. J Appl Environ Microbiol. 2015; 81(15): 5196–5202.

https://doi.org/10.1128/aem.00832-15

[22] Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020; 578: 432–436. https://doi.org/10.1038/s41586-020-1957-x

[23] Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol. 2019; 4: 1895–1906. https://doi.org/10.1038/s41564-019-0510-x

[24] Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C, Expansion of known ssRNA phage genomes: From tens to over a thousand. Sci Adv. 2020: 6(6): E5981. https://doi.org/10.1093/nar/gkw1030

[25] Paez-Espino D, Chen IMA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2016; 45(1): 457–465. https://doi.org/10.1038/nature19094

[26] Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016; 536: 425–430. https://doi.org/10.1038/nature14486

[27] Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015; 523: 208–211. https://doi.org/10.1038/ismej.2012.59

[28] Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J. 2012; 6: 1715–1727.

https://doi.org/10.1038/ismej.2012.78

[29] Marshall K, Morris R. Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. ISME J. 2013; 7: 452–455. https://doi.org/10.1038/nrgastro.2011.191

[30] Man SM. The clinical importance of emerging Campylobacter species. Nat. Rev. Gastroenterol Hepatol. 2011; 8(12): 669–685. https://doi.org/10.1016/S1473-3099(10)70251-6

[31] Christensen H, May M, Bowen L, Hickman M, Trotter CL. Meningococcal carriage by age: a systematic review and meta-analysis. Lancet Infect. Dis. 2010; 10: 853–861. https://doi.org/10.1038/s41579-019-0282-6

[32] Caugant DA, Brynildsrud OB. Neisseria meningitidis: using genomics to understand diversity, evolution and pathogenesis. Nat Rev Microbiol. 2020; 18: 84–96. https://doi.org/10.1128/JB.183.8.2570-2575.2001

[33] Claus H, Stoevesandt J, Frosch M, Vogel U. Genetic isolation of meningococci of the electrophoretic type 37 complex. J Bacteriol. 2001; 183 (8): 2570-2575. https://doi.org/10.1038/s41564-018-0166-y

[34] Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018; 3:754–766. https://doi.org/10.1128/jb.171.12.6845-6849.1989

[35] Toro N, Datta A, Carmi OA, Young C, Prusti RK, Nester EW. The Agrobacterium tumefaciens virC1 gene product binds to overdrive, a T-DNA transfer enhancer. J Bacteriol. 1989; 171 (12): 6845-6849.

https://doi.org/10.1146/annurev.micro.59.030804.121320

[36] iboud GI, Bliska JB. Yersinia outer proteins: role in modulation of host cell signaling responses and pathogenesis. Annu Rev Microbiol. 2005; 59:69-89. https://doi.org/10.1016/j.tim.2006.04.005

[37] Linke D, Riess T, Autenrieth IB, Lupas A, Kempf VA. Trimeric autotransporter adhesins: variable structure, common function. Trends Microbiol. 2006; 14(6):264-70. https://doi.org/10.1006/bbrc.2001.5591

[38] Chang KW, Weng SF, Tseng YH. UDP-glucose dehydrogenase gene of Xanthomonas campestris is required for virulence. Biochem Biophys Res Commun. 2001; 287(2):550-555. https://doi.org/10.1021/bi025862m

[39] Naught LE, Gilbert S, Imhoff R, Snook C, Beamer L, Tipton P. Allosterism and cooperativity in Pseudomonas aeruginosa GDP-mannose dehydrogenase. Biochemistry. 2002; 41(30):9637-9645. https://doi.org/10.1128/JB.181.1.141-148.1999

[40] Núñez C, Moreno S, Soberón-Chávez G, Espín G. The Azotobacter vinelandii response regulator AlgR is essential for cyst formation. J Bacteriol. 1999;181(1):141-148. https://doi.org/10.1074/jbc.274.16.10936

[41] Bahassi EM, O'Dea MH, Allali N, Messens J, Gellert M, Couturier M. Interactions of CcdB with DNA gyrase. Inactivation of Gyra, poisoning of the gyrase-DNA complex, and the antidote action of CcdA. J Biol Chem. 1999; 274(16):10936-10944. https://doi.org/10.1073/pnas.85.9.2909

[42] Das A. Agrobacterium tumefaciens virE operon encodes a single-stranded DNA-

binding protein. Proc Natl Acad Sci U S A. 1988; 85(9):2909-2913. https://doi.org/10.1016/j.micinf.2007.01.020

[43] Selvaraj SK, Periandythevar P, Prasadarao NV. Outer membrane protein A of Escherichia coli K1 selectively enhances the expression of intercellular adhesion molecule-1 in brain microvascular endothelial cells. Microbes Infect. 2007; 9(5):547-557. https://doi.org/10.1038/nrmicro2651

[44] Yamaguchi Y, Inouye M. Regulation of growth and death in Escherichia coli by toxin-antitoxin systems. Nat Rev Microbiol. 2011; 9(11):779-790. https://doi.org/10.1073/pnas.0434325100

[45] Meinhart A, Alonso JC, Sträter N, Saenger W. Crystal structure of the plasmid maintenance system epsilon/zeta: functional mechanism of toxin zeta and inactivation by epsilon 2 zeta 2 complex formation. Proc Natl Acad Sci U S A. 2003;100(4):1661-1666. https://doi.org/10.1099/00221287-146-7-1535

[46] McCrea KW, Hartford O, Davis S, Eidhin DN, Lina G, Speziale P, et al. The serine-aspartate repeat (Sdr) protein family in Staphylococcus epidermidis. Microbiology. 2000; 146: 7. https://doi.org/10.1016/s0969-2126(01)00255-6

[47] Gaskell S, Crennell S, Taylor G. The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure. 1995; 3(11):1197-1205. https://doi.org/10.1073/pnas.0506758102

[48] Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome.” Proc Natl Acad Sci U S A. 2005;

102(39): 13950–13955. https://doi.org/10.1016/S1367-5931(99)00013-7

[49] Schmidt M, Lupas AN, Finley D, Structure and mechanism of ATP-dependent proteases. Curr Opin Chem Biol. 1999; 3: 584–591. https://doi.org/10.1146/annurev-biochem-060408-172623

[50] Sauer RT, Baker TA. AAA+ proteases: ATP-fueled machines of protein destruction. Annu Rev Biochem. 2011; 80: 587–612. https://doi.org/10.1016/j.gde.2005.09.006

[51] D. Medini, C. Donati, H. Tettelin, V. Masignani, R. Rappuoli, The microbial pan-genome. Curr. Opin. Genet. Dev. 15(6), 589–594 (2005). https://doi.org/10.1006/jtbi.2002.3054

[52] Krakauer DC, Jansen VA. Red queen dynamics of protein translation. J Theor Biol. 2002; 218: 97–109. https://doi.org/10.1038/s41559-020-1124-7

[53] Chen F, Wu P, Deng S, Zhang H, Hou Y, Hu Z, et al. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. Nat Ecol Evol. 2020; 4: 589–600. https://doi.org/10.1038/ismej.2011.146

[54] Enav H, Béjà O, Mandel-Gutfreund Y. Cyanophage tRNAs may have a role in cross-infectivity of oceanic Prochlorococcus and Synechococcus hosts. ISME J. 2011; 6(3): 619–628. https://doi.org/10.1111/1462-2920.14326

[55] Xu Y, Zhang R, Wang N, Cai L, Tong Y, Sun Q, et al. Novel phage-host interactions and evolution as revealed by a cyanomyovirus isolated from an estuarine environment. Environ Microbiol. 2018; 20: 2974-2989. https://doi.org/10.1093/nar/29.5.1097

[56] Sofia HJ, Chen G, Hetzler BG, Reyes-Spindola JF, Miller NE. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. Nucleic Acids Res. 2001; 29:1097-1106. https://doi.org/10.1073/pnas.0504068102

[57] Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A. 2005; 102: 14332–14337. https://doi.org/10.1101/gr.5322306

[58] Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 2006; 16: 1099–1108. https://doi.org/10.1073/pnas.1001418107

[59] Andam CP, Williams D, Gogarten JP. Biased gene transfer mimics patterns created through shared ancestry. Proc Natl Acad Sci U S A. 2010; 107: 10679–10684. https://doi.org/10.1007/s11104-008-9668-3

[60] Herridge D, Peoples M, Boddey R. Global inputs of biological nitrogen fixation in agricultural systems. Plant Soil. 2008; 311: 1–18. https://doi.org/10.1038/nrmicro.2017.171

[61] Poole P, Ramachandran V, Terpolilli J. Rhizobia: from saprophytes to endosymbionts. Nat Rev Microbiol. 2018; 16(5): 291–303. https://doi.org/10.1038/nrmicro2910

[62] Vorholt J. Microbial life in the phyllosphere. Nat Rev Microbiol. 2012; 10: 828–

840. https://doi.org/10.1046/j.1364-3703.2003.00149.x.

[63] Hingamp P, Grimsley N, Acinas S, Clerissi C, Subirana L, Poulain J, et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J. 2013; 7: 1678–1695. https://doi.org/10.1038/ismej.2013.59

[64] Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics. 2016; 32: 1933–1942. https://doi.org/10.1093/bioinformatics/btw108

[65] Arndt W. Modifying HMMER3 to run efficiently on the Cori supercomputer using OpenMP tasking. In Proc. 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). 2018; 239–246.

[66] 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. https://doi.org/10.1007/978-1-4899-7478-5_221

[67] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015: 12: 59–60. https://doi.org/10.15496/publikation-1176

[68] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families' database in 2019. Nucleic Acids Res. 2019; 47(1): 427–432. https://doi.org/10.1093/nar/gky995

[69] Lowe TM, Chan PP. tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Nucleic Acids Res. 2016; 44: 54-57. https://doi.org/10.1093/nar/gkw413

[70] Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25: 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

[71] Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE. 2016; 11(10): e0163962. https://doi.org/10.1371/journal.pone.0163962

[72] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 2020; 37:1530-1534. https://doi.org/10.1101/849372

[73] Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016; 44: 242–245. https://doi.org/10.1093/nar/gkw290

[74] Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16:157. https://doi.org/10.1186/s13059-015-0721-2

[75] Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. Bioinformatics. 2017; 33: 2379-2380. https://doi.org/10.1093/bioinformatics/btx157

[76] Wei S. Jie X. TaxonKit - A Cross-platform and Efficient NCBI Taxonomy Toolkit. https://www.biorxiv.org/content/10.1101/513523v1 (2019). https://doi.org/10.1101/513523

[77] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In Proceeding International AAAI Conference on Weblogs and Social Media, San Jose, CA. 2009: 17-20.

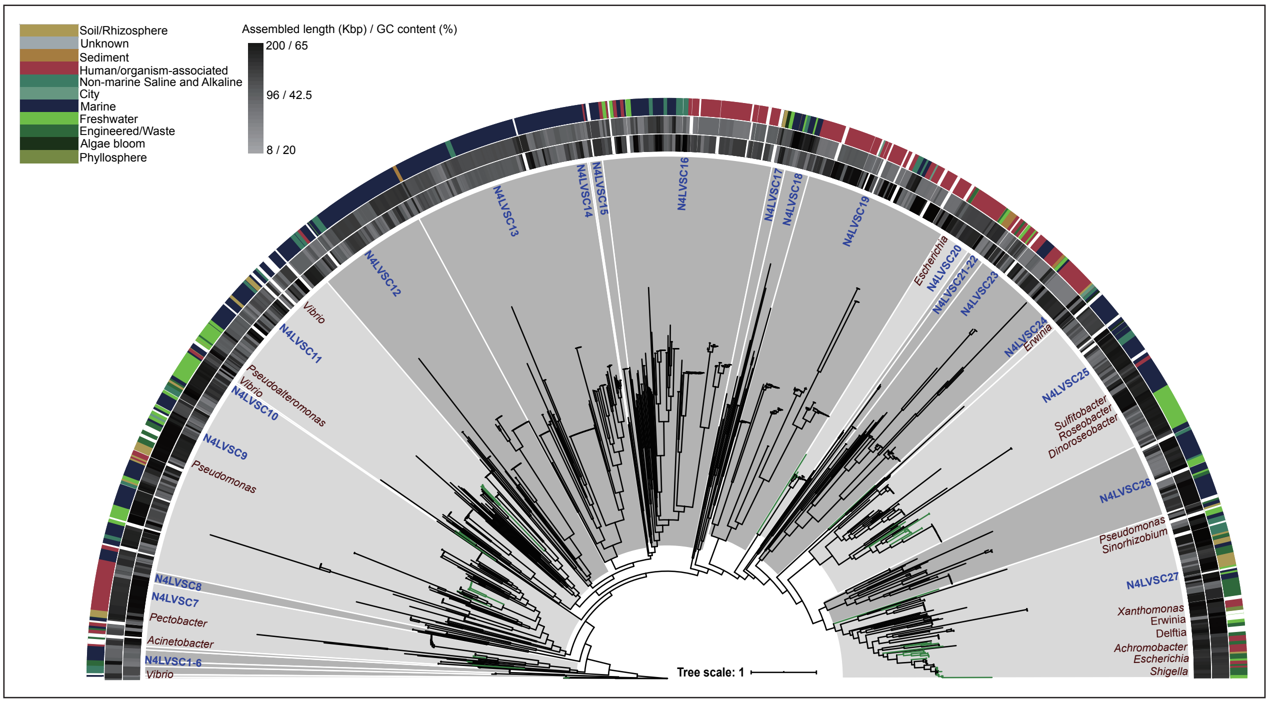


Fig. 1 The phylogenic tree of maximum likelihood was generated from concatenated two core-gene (N4-like major capsid protein and N4-like virion-encapsulated RNA polymerase) displayed the diverse subclades of N4-like virus (N4LVSC). Total 27 N4LVSCs assigned in phylogenic tree, sub-clades are colored by light grey or dark grey, presenting isolated virus existent or without existent respectively, and the isolates are labeled as green bold branches. The annotations of the tree from outside to inner side is origin of corresponding viral contig (color legend is showed on the upper left side), G+C content, assemble length and host names of isolated N4-like viruses (N4LVs) (colored by burgundy)/number of N4LVSCs (colored by skyblue) separately.

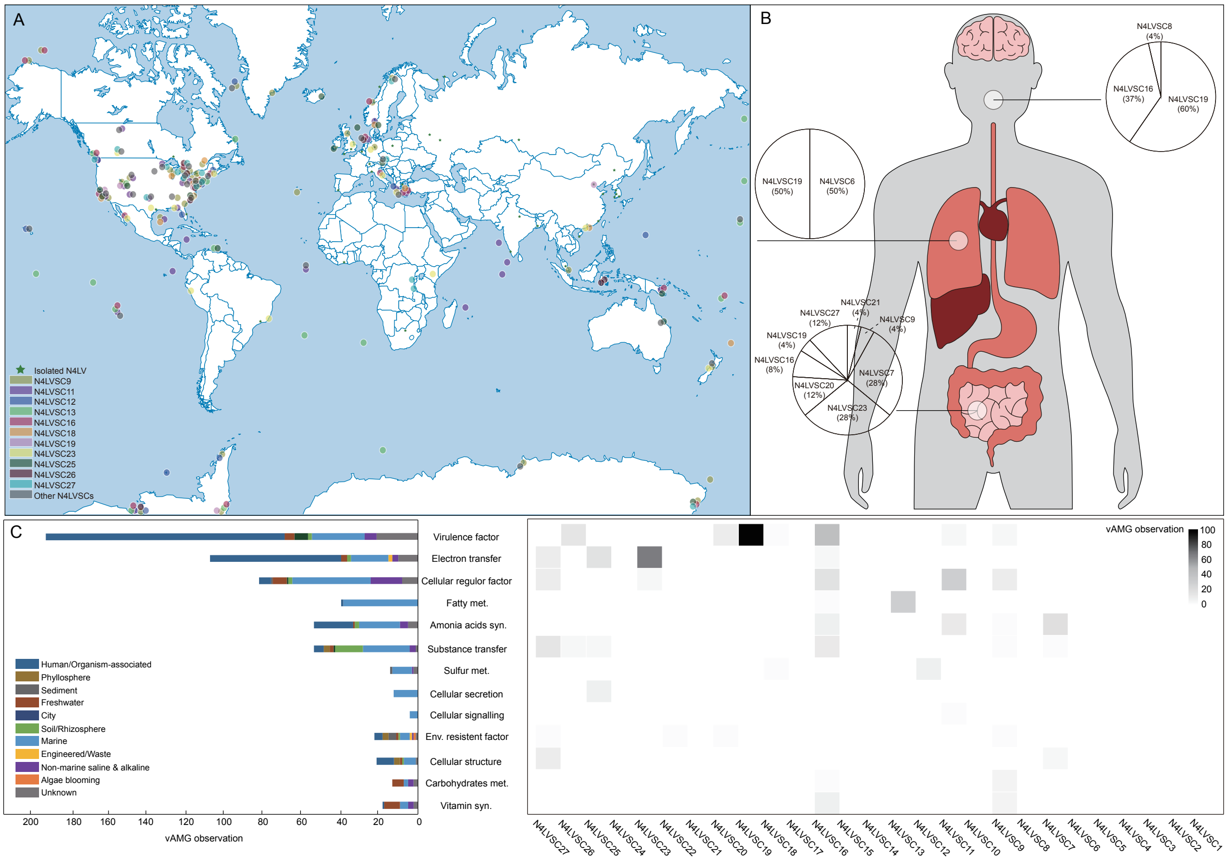
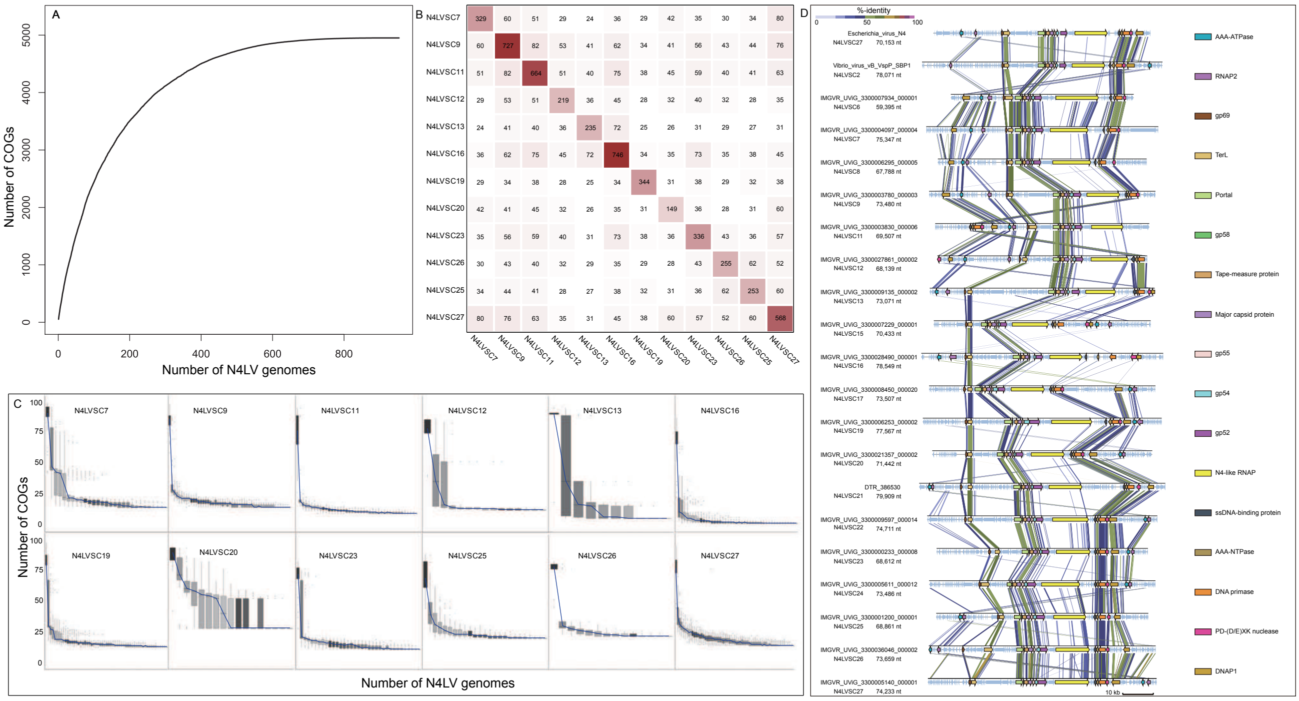


Fig. 2. The source of N4LV in global scale and human body respectively, as well as the distribution of potential viral-encoded auxiliary metabolic genes (vAMGs). (A) The origins of N4LVSCs are showed in the map in the accordance of information providing by IMG/M. Isolated N4LVs are labeled as green star, while other metagenomic assembly genomes (N4LVMAGs) are labeled as the colored nodes. Top 11 N4LVSCs containing over 90% N4LVs are labeled as specific colors, while other 16 N4LVSCs are labeled as grey. (B) The human-associated N4LVSCs are showed in the diagram of human body in the accordance of information providing by IMG/M. Four main viral metagenomics sources are referred to human body, including oral cavity, skin, respiratory system and guts. The components of different N4LVSCs in oral cavity are showed in the pie chart. (C) The observation of potential vAMGs was classified as 13 groups. The bar chart on the bottom left side indicates distribution in diverse habitats, and the heatmap on the bottom right side indicates distribution in different N4LVSCs.



	N4LVSC7	N4LVSC9	N4LVSC11	N4LVSC12	N4LVSC13	N4LVSC16	N4LVSC19	N4LVSC20	N4LVSC23	N4LVSC26	N4LVSC25	N4LVSC27
N4LVSC7	329	60	51	29	24	36	29	42	35	30	34	80
N4LVSC9	60	727	82	53	41	62	34	41	56	43	44	76
N4LVSC11	51	82	664	51	40	75	38	45	59	40	41	63
N4LVSC12	29	53	51	219	36	45	28	32	40	32	28	35
N4LVSC13	24	41	40	36	235	72	25	26	31	29	27	31
N4LVSC16	36	62	75	45	72	746	34	35	73	35	38	45
N4LVSC19	29	34	38	28	25	34	344	31	38	29	32	38
N4LVSC20	42	41	45	32	26	35	31	149	36	28	31	60
N4LVSC23	35	56	59	40	31	73	38	36	336	43	36	57
N4LVSC26	30	43	40	32	29	35	29	28	43	255	62	52
N4LVSC25	34	44	41	28	27	38	32	31	36	62	253	60
N4LVSC27	80	76	63	35	31	45	38	60	57	52	60	568

Fig. 3 The pan-/core-genome analysis of N4LVs. (A) Pangenome curve was plotted based on the protein groups of orthologs (N4LVOGs) accumulation with all 1000 N4LV genomes accumulation. Medians of each sampling under complete random were linked to form this curve. (B) The number of shared N4LVOGs between N4LVSCs is showed in the heatmap. The deeper the color, the more N4LVOGs are included in this N4LVSC, and the number of N4LVOGs is displayed inside each box of the heatmap. Only high-quality N4LVs in at least 10 members-containing N4LVSC were included in this analysis to avoid bias of incomplete genomic fragment or insufficient sample size. (C) The core-genome curve with sampling boxes plot were plotted based on core N4LVOG dilution with the high-quality N4LV genomes accumulation. (D) Genomic synteny analysis of 21 complete or nearly complete N4LVs. Conserved proteins in genomes are labeled as different colors. There were 21 high-quality genomes of N4LV selected to map synteny of comparative genomics.

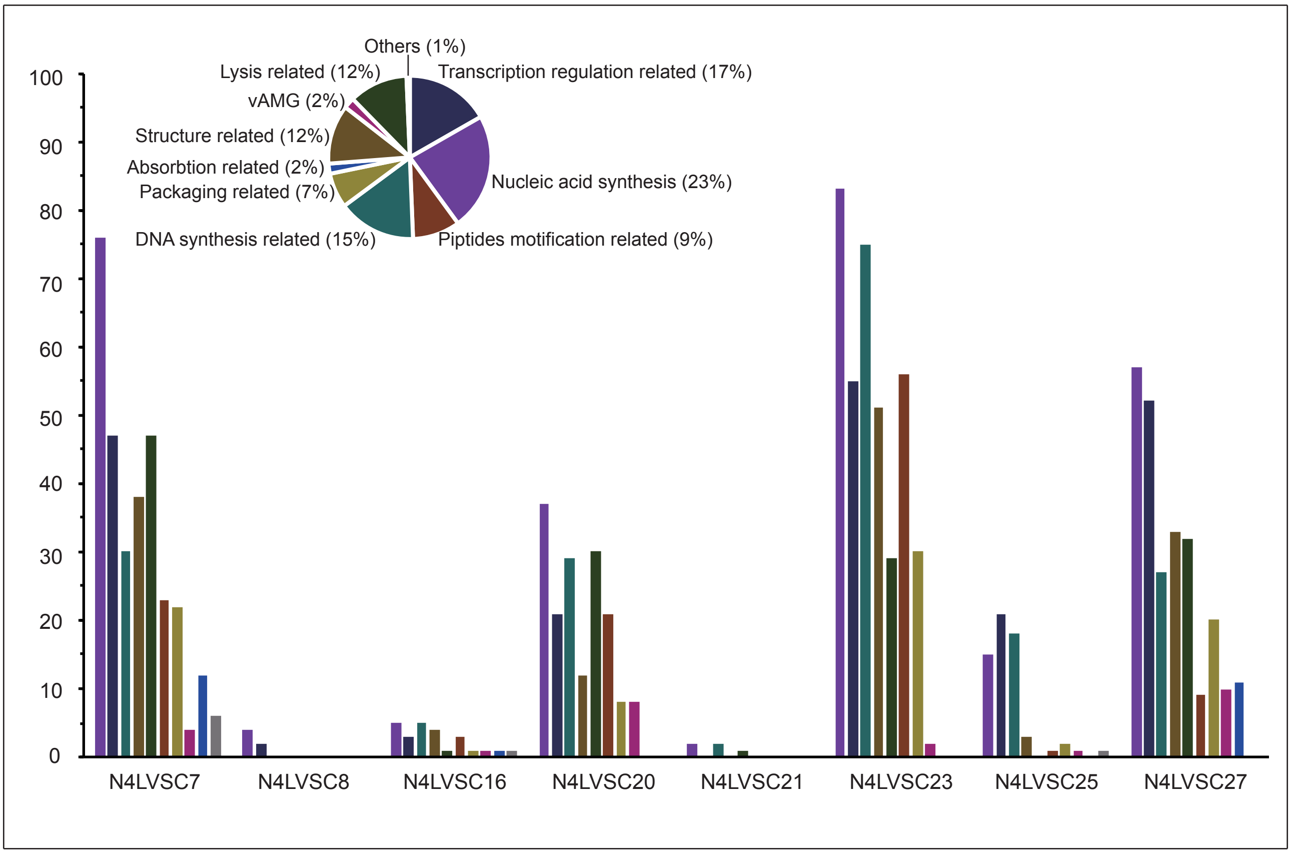


Fig. 4 Number of transcriptions compensated genes by viral tRNA in different N4LVSCs. The genes included in Bin 2 and Bin 3 were classified as 10 types, indicating by the pie chart in different colors. The observations of diverse types of gene are varied in different N4LVSCs, which is indicated by the bar chart. The percentage in the pie chart shows the ratio of each component comparing with the whole.

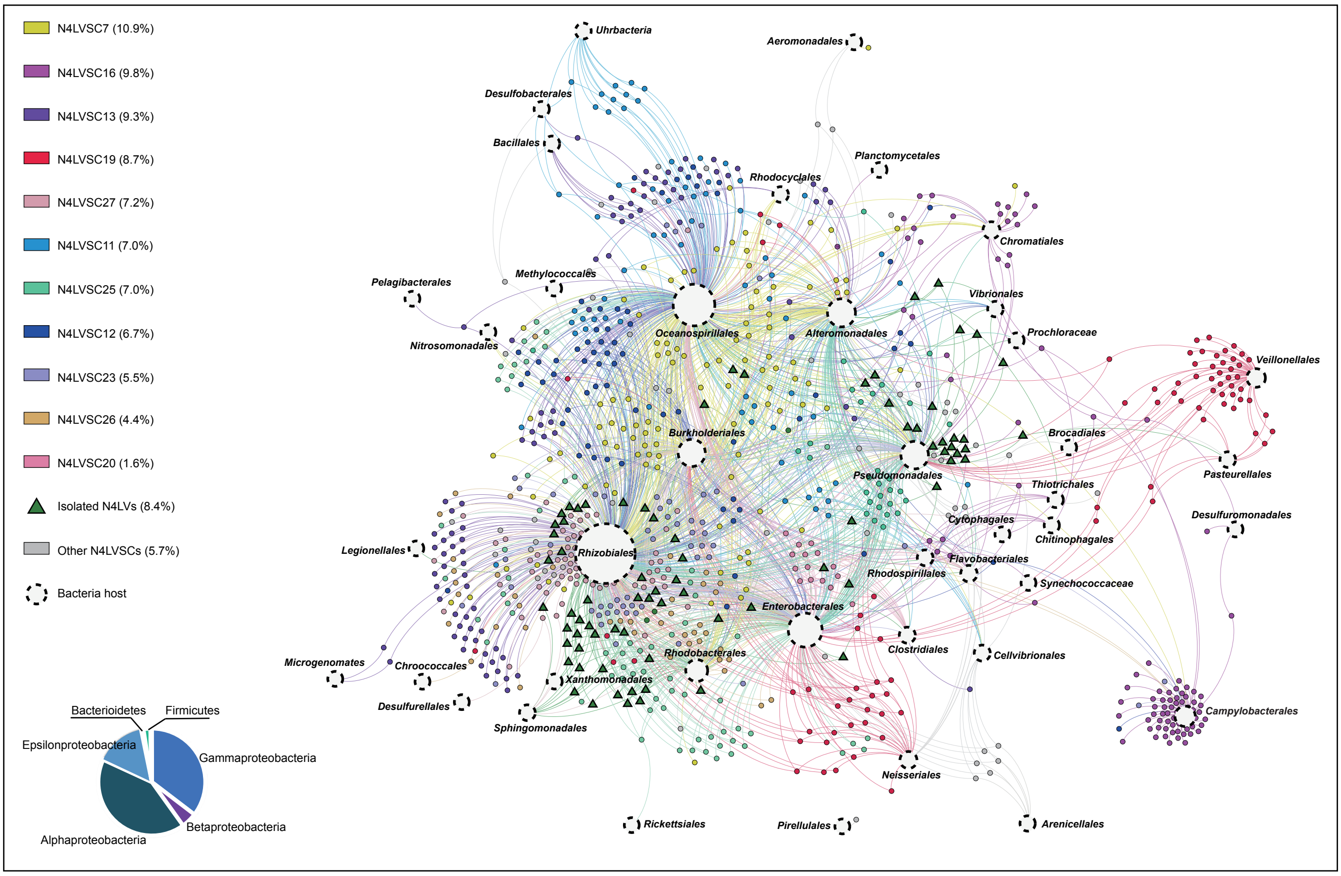


Fig. 5 The horizontal gene transfer (HGT) linkage network of high-quality N4LVs with their putative hosts. Top 11 N4LVSCs that occupied over 90% HGT linkages are labeled as specific colors, while other 15 N4LVSCs, isolated N4LVs are labeled as grey dots and green triangles, respectively. The HGT-occurred bacteria hosts are labeled as hollow circles with dash line, which names are labeled near corresponding hollow circles. The pie chart on the bottom left illustrated the percentage of HGT linkage observation for each taxonomic category in phylum level

Figures

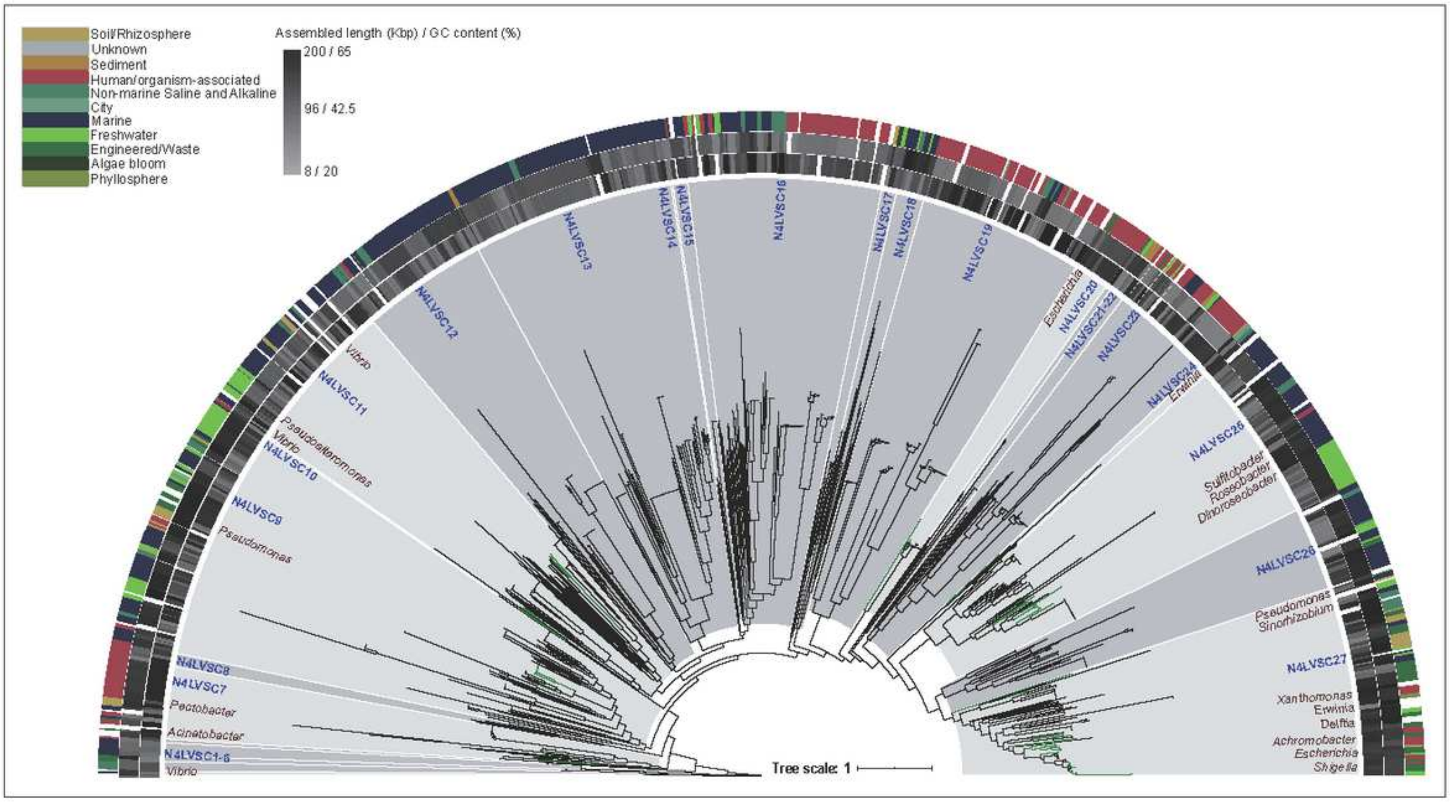


Figure 1

The phylogenic tree of maximum likelihood was generated from concatenated two core-gene (N4-like major capsid protein and N4-like virion-encapsulated RNA polymerase) displayed the diverse subclades of N4-like virus (N4LVSC). Total 27 N4LVSCs assigned in phylogenic tree, sub-clades are colored by light grey or dark grey, presenting isolated virus existent or without existent respectively, and the isolates are labeled as green bold branches. The annotations of the tree from outside to inner side is origin of corresponding viral contig (color legend is showed on the upper left side), G+C content, assemble length and host names of isolated N4-like viruses (N4LVs) (colored by burgundy)/number of N4LVSCs (colored by skyblue) separately.

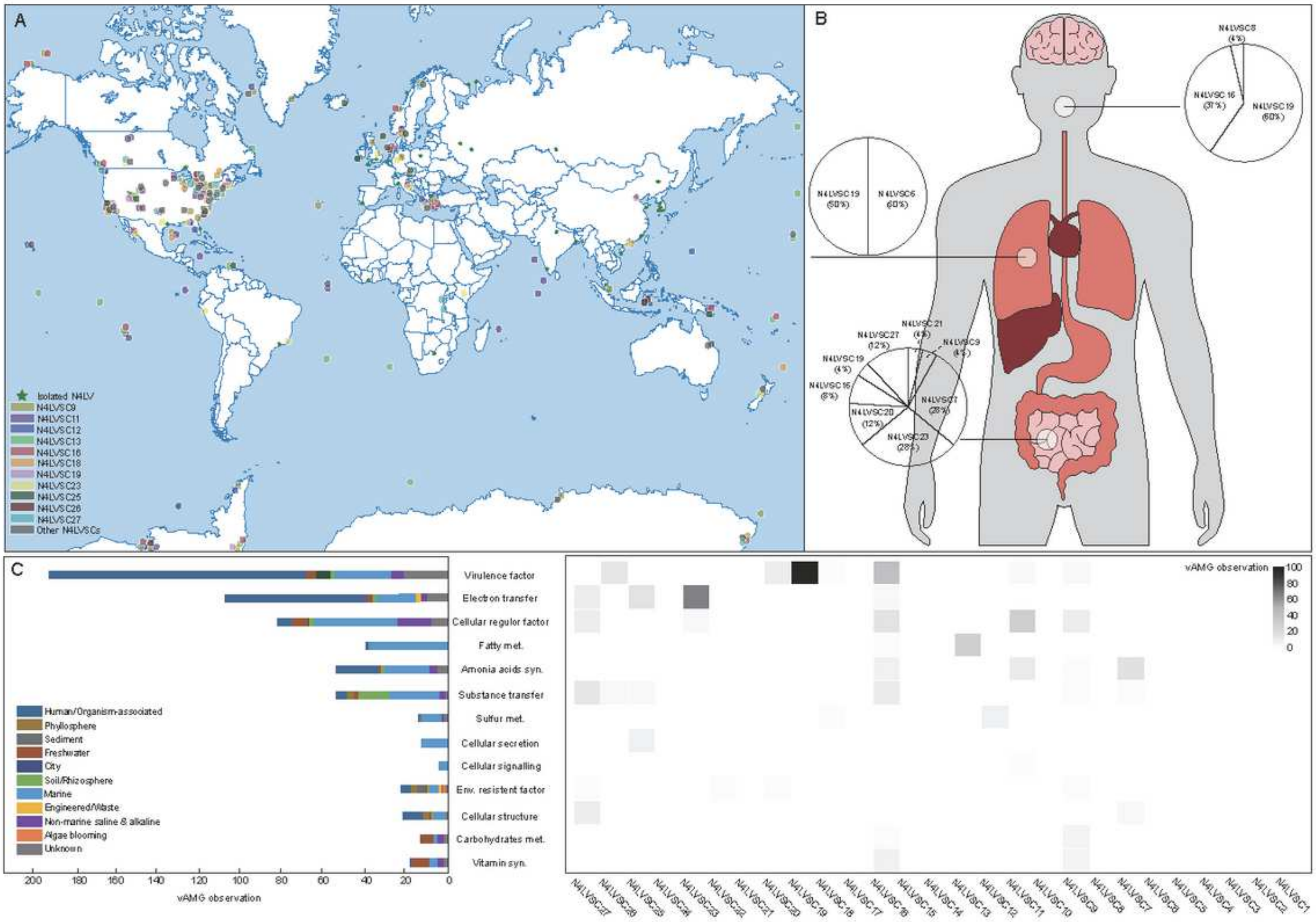


Figure 2

The source of N4LV in global scale and human body respectively, as well as the distribution of potential viral-encoded auxiliary metabolic genes (vAMGs). (A) The origins of N4LVSCs are showed in the map in the accordance of information providing by IMG/M. Isolated N4LVs are labeled as green star, while other metagenomic assembly genomes (N4LVMAGs) are labeled as the colored nodes. Top 11 N4LVSCs containing over 90% N4LVs are labeled as specific colors, while other 16 N4LVSCs are labeled as grey. (B) The human-associated N4LVSCs are showed in the diagram of human body in the accordance of information providing by IMG/M. Four main viral metagenomics sources are referred to human body, including oral cavity, skin, respiratory system and guts. The components of different N4LVSCs in oral cavity are showed in the pie chart. (C) The observation of potential vAMGs was classified as 13 groups. The bar chart on the bottom left side indicates distribution in diverse habitats, and the heatmap on the bottom right side indicates distribution in different N4LVSCs. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area o bbnhjr of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

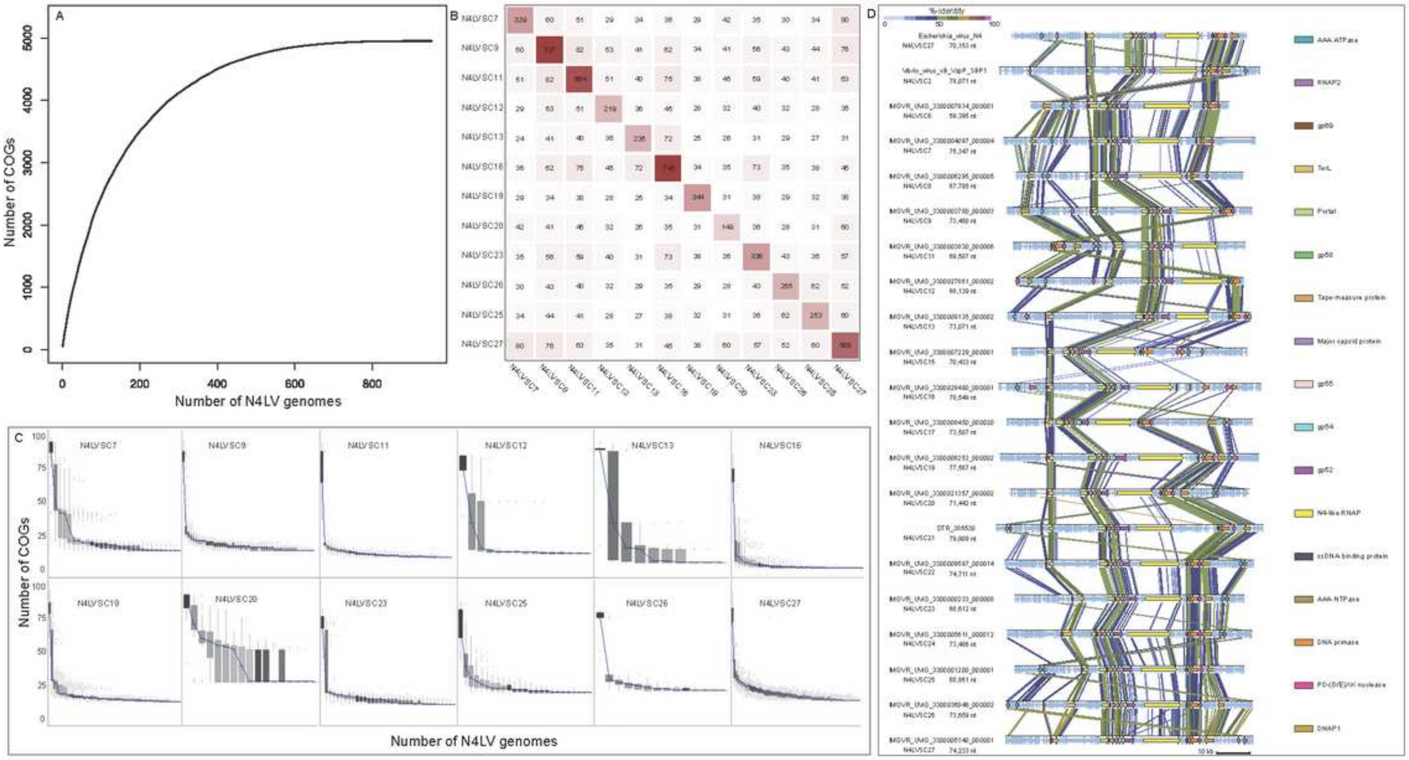


Figure 3

The pan-/core-genome analysis of N4LVs. (A) Pangenome curve was plotted based on the protein groups of orthologs (N4LVOGs) accumulation with all 1000 N4LV genomes accumulation. Medians of each sampling under complete random were linked to form this curve. (B) The number of shared N4LVOGs between N4LVSCs is showed in the heatmap. The deeper the color, the more N4LVOGs are included in this N4LVSC, and the number of N4LVOGs is displayed inside each box of the heatmap. Only high-quality N4LVs in at least 10 members-containing N4LVSC were included in this analysis to avoid bias of incomplete genomic fragment or insufficient sample size. (C) The core-genome curve with sampling boxes plot were plotted based on core N4LVOG dilution with the high-quality N4LV genomes accumulation. (D) Genomic synteny analysis of 21 complete or nearly complete N4LVs. Conserved proteins in genomes are labeled as different colors. There were 21 highquality genomes of N4LV selected to map synteny of comparative genomics.

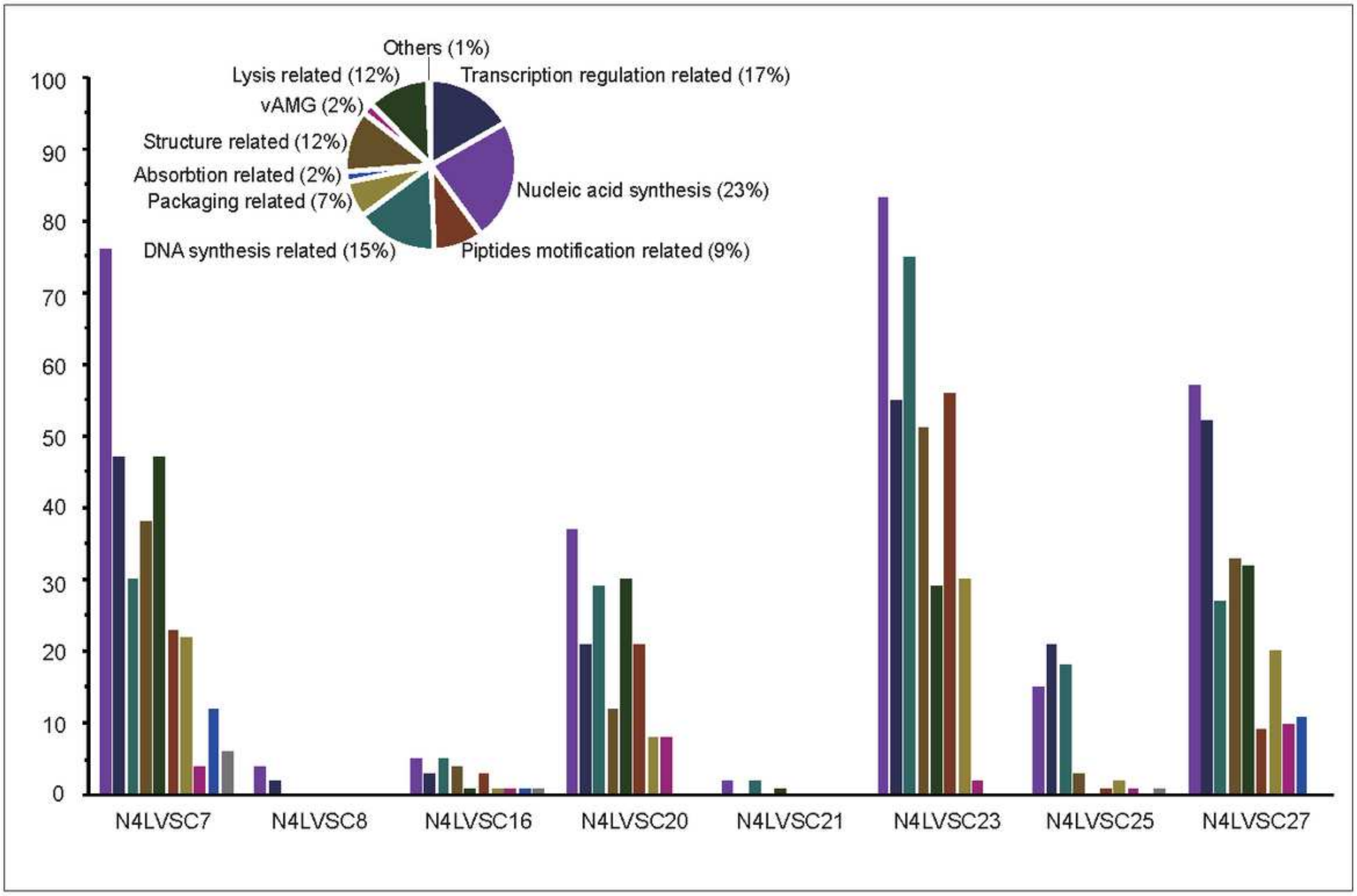


Figure 4

Number of transcriptions compensated genes by viral tRNA in different N4LVSCs. The genes included in Bin 2 and Bin 3 were classified as 10 types, indicating by the pie chart in different colors. The observations of diverse types of gene are varied in different N4LVSCs, which is indicated by the bar chart. The percentage in the pie chart shows the ratio of each component comparing with the whole.

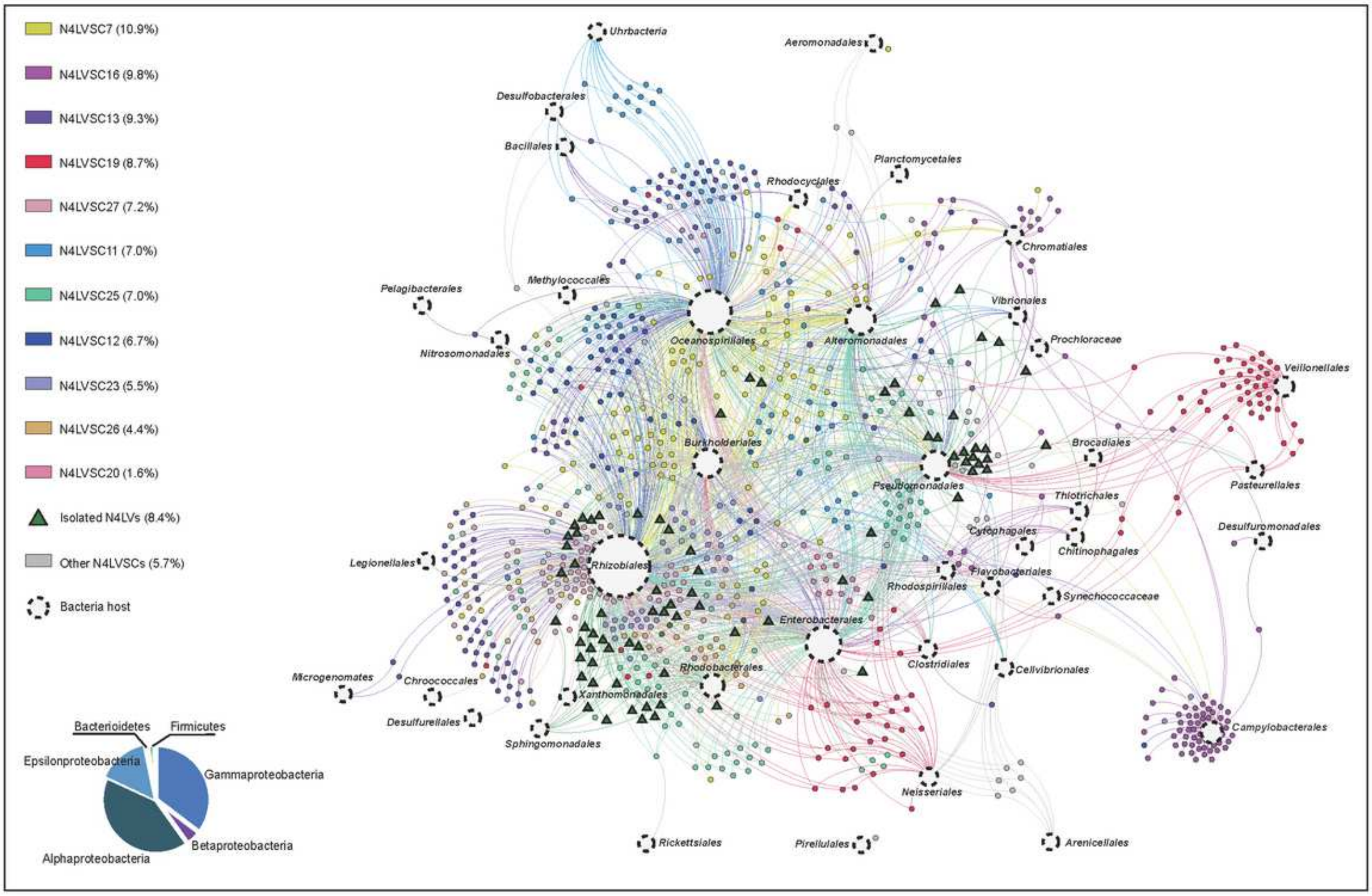


Figure 5

The horizontal gene transfer (HGT) linkage network of high-quality N4LVs with their putative hosts. Top 11 N4LVSCs that occupied over 90% HGT linkages are labeled as specific colors, while other 15 N4LVSCs, isolated N4LVs are labeled as grey dots and green triangles, respectively. The HGT-occurred bacteria hosts are labeled as hollow circles with dash line, which names are labeled near corresponding hollow circles. The pie chart on the bottom left illustrated the percentage of HGT linkage observation for each taxonomic category in phylum level

Supplementary Files

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

- SupplementaryData1.tsv
- SupplementaryData2.tsv
- SupplementaryData3.tsv
- SupplementaryData45.zip
- SupplementaryMaterials.pdf