# Temporal dynamics of uncultured viruses: a new dimension in viral diversity

Arkhipova, K., Skvortsov, T., Quinn, J., McGrath, J., Allen, C., Dutilh, B., ... Kulakov, L. (2017). Temporal dynamics of uncultured viruses: a new dimension in viral diversity. DOI: 10.1038/ismej.2017.157

**Published in:**
The ISME Journal

**Document Version:**
Peer reviewed version

**Title: Temporal dynamics of uncultured viruses: a new dimension in viral diversity**

**Authors:** Ksenia Arkhipova[1,2], Timofey Skvortsov [1,3], John P. Quinn[1], John W. McGrath[1,3], Christopher C.R. Allen[1,3], Bas E. Dutilh[2,4], Yvonne McElarney[5], Leonid A. Kulakov[1*]

**Affiliations:**

[1]School of Biological Sciences, The Queen's University of Belfast, Belfast, Northern Ireland, UK

[2]Theoretical Biology and Bioinformatics, Utrecht University, 3584 CH, Utrecht, the Netherlands

[3]Institute for Global Food Security, The Queen's University of Belfast, Belfast, Northern Ireland, UK

[4]Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, 6525 GA, Nijmegen, the Netherlands

[5]Agri-Food and Biosciences Institute, Belfast, Northern Ireland, UK

[*]Correspondence to l.kulakov@qub.ac.uk

**Abstract:**

Recent work has vastly expanded the known viral genomic sequence space, but the seasonal dynamics of viral populations at the genome level remain unexplored. Here we followed the viral community in a freshwater lake for one year using genome-resolved viral metagenomics, combined with detailed analyses of the viral community structure, associated bacterial populations, and environmental variables. We reconstructed 8,950 complete and partial viral genomes, the majority of which were not persistent in the lake throughout the year, but instead continuously succeeded each other. Temporal analysis of 732 viral genus-

level clusters demonstrated that one fifth were undetectable at specific periods of the year. Based on host predictions for a subset of reconstructed viral genomes, we for the first time reveal three distinct patterns of host-pathogen dynamics, where the viruses may peak before, during, or after the peak in their host's abundance, providing new possibilities for modelling of their interactions. Time-series metagenomics opens up a new dimension in viral profiling, which is essential to understand the full scale of viral diversity and evolution, and the ecological roles of these important players in the global ecosystem.

**Main text:**

One of the major challenges in studies of viral dynamics is the absence of a phylogenetically informative universal marker, analogous to the bacterial 16S or eukaryotic 18S rRNA genes. To analyse temporal changes of some viral subgroups (e.g. marine T4-like myoviruses or freshwater cyanomyoviruses), recent studies have used sequencing of amplicons of viral conserved structural proteins, such as capsid proteins g23 or g20 (Chow & Fuhrman, 2012; Yeo & Gin, 2015; Wang et al, 2015). However, this approach does not allow assessment of the dynamics of the whole community. A shotgun metagenomics approach does not share this limitation and provides a means to study seasonal changes without any *a priori* assumptions about the structure of a viral community. Using shotgun metagenomics, some attempts have been made to study viral dynamics, for example by tracking the temporal changes of 35 individual de novo assembled viral genomes (Emerson et al, 2012), or by binning sequencing reads into assemblages (possibly at a viral family level (Bolduc et al, 2015)) to study their temporal stability and/or fluctuations (Bolduc et al, 2015; Emerson et al, 2013). Although these studies have provided much-needed insight into possible scenarios of viral dynamics, there is still no global picture available of seasonal changes of viral populations and their links to other factors in an ecosystem.

Due to the mosaic nature of viral genome organisation, assessment of viral genetic similarity is a non-trivial task. To tackle this problem, Lima-Mendez et al. in 2008 proposed a method of reticulate classification of phage genetic relatedness (Lima-Mendez et al, 2008). The method provides means to subdivide the whole sequence space of viral metagenomics data into groups approximately corresponding to genus level of taxonomical classification. Since that time the approach has been successfully used in several studies to gain deeper insight into phage biology and to connect newly assembled genomes with already known sequences (Roux et al, 2016; Roux et al, 2015). At the same time, it is well known that sequence relatedness within characterised viral genera can vary substantially (King et al, 2011), but in natural environments the genetic variation of newly assembled viral genomes within 'genera' resulting from reticulate clustering has not yet been analysed.

Along with the gaps in knowledge of global viral sequence diversity, there is a lack of information about the possible variants of bacteria-phage dynamic interactions. To date, a range of models describing behaviour of some host-pathogen relationships have been developed. First and foremost, the Kill-the-Winner model (Thingstad, 2000), which assesses populations' changes within the framework of the classic Lotka-Volterra model. Recently, Knowles et al. have noticed discrepancies between the predictions of the model and the experimentally measured virus and host abundances in natural environments (Knowles et al, 2016), which poses a question about the possible existence of other dynamics of host-pathogen interactions in natural microbial communities.

Here we present a detailed exploration of the structure, seasonal dynamics and functional potential of the viral community in a temperate freshwater eutrophic lake (Lough Neagh, Northern Ireland). Our novel data includes 12 viral shotgun metagenomes and 13 bacterial 16S rRNA-amplicon datasets collected over a period of one year (Supplementary Table 1, Sheet 1). This unique collection of data allowed us to explore the range of interaction

dynamics of viruses and their hosts in a natural ecosystem. We also investigate the possibility of functional manipulations of bacteria by phages by analysing auxiliary metabolic genes, revealing that their functions are clearly different in winter compared to summer.

## Material and Methods

### Data availability

Raw reads from the Illumina sequencing and sequences of bacterial 16S rRNA gene amplicons are available for download from the Short Reads Archive (BioProject PRJNA350258 and PRJNA292054). Annotated viral reads and assembled sequences are also available on MetaVir and MG-RAST databases (for accession numbers see Supplementary Table 1, Sheet 1).

### Sample collection, processing and sequencing

Lough Neagh is a large eutrophic polymictic shallow freshwater lake located in Northern Ireland (UK). Water samples were collected from the deepest site in the lake (54°37′06″N, 6°23′43″W) at 12 time points over the period of a year (Supplementary Table 1, Sheet 1) as described previously (Skvortsov et al, 2016). Some environmental parameters, such as temperature and pH at 5 m depth were recorded at the collection site and several extra water samples were taken for chemical analysis (Supplementary Table 5, Sheet 2). Sample processing steps, DNA extraction, library preparation and sequencing procedures have been described in detail previously (Skvortsov et al, 2016). Briefly, water samples were filtered through 0.22 μm filters to obtain a 'virus-like particle' (VLP) water fraction, which was concentrated using 100 kDa filters and treated with DNAse I. Extracted and purified DNA was used for library preparation with Nextera DNA Sample Preparation kit (Illumina, USA) and sequenced from both ends with the 600-cycle MiSeq Reagent Kit v3 on MiSeq (Illumina, USA) at the University of Cambridge DNA Sequencing facility.

Total DNA (particle sizes more than 0.22 μm) was extracted from 500 ml of water using a PowerWater DNA Isolation kit (MO BIO, USA). Partial bacterial 16S rRNA gene sequences were amplified with 909-F/1492-R primers and sequenced on a 454 GS Junior (Roche, USA) with Lib-L Shotgun chemistry.

Sequencing library processing, assembly and annotation

The Illumina reads were processed with BBMap v 33.54 (http://sourceforge.net/projects/bbmap/) software, and all reads with an average Q-score < 15 or containing Ns were discarded. We applied a two-step assembly strategy. First all 12 libraries were assembled separately using the graph-based assembler IDBA-UD (Peng et al, 2012) (kmer range 20-250, step - 10). Next, all the libraries were combined and assembled collectively (kmer range 20-1500, step - 10). This allowed us to use all available reads in the assembly to reconstruct even low-abundance viral genomes as well as to maximise assembly effectiveness for genomes appearing only in individual libraries. After that, an additional attempt to elongate the contigs obtained in the two previous steps was made using an overlap-layout-consensus assembler with very strict parameters (CAP3 (Huang & Madan, 1999), overlap > 2,000bp, percentage of nucleotide identity - 99%). This step also reduced drastically the number of duplicated sequences. To completely remove duplicates and leave only the longest assembled contigs, we used the cd-hit (Li & Godzik, 2006) program (-c 0.98 -n 11 -d 0). For subsequent analyses only sequences longer than 7,000bp were retained. To estimate what part of the viral population this set of contigs represented, reads from all 12 libraries were mapped onto contigs using BBMap (70% of nucleotide identity).

Open reading frames in the assembled contigs were predicted with MetaGeneAnnotator (Noguchi et al, 2008). For functional annotation, the contigs assembled separately from 12 libraries were uploaded to the MG-RAST (Meyer et al, 2008) and MetaVir (Roux et al, 2014) servers (please see Supplementary Table 1, Sheet 1 for the accession numbers). The resulting

functional annotations with SEED Subsystems were downloaded from MG-RAST, percentages of all categories were calculated for each individually annotated library and were used in correlation analysis.

Raw reads obtained from the sequencing of 16S rRNA gene amplicons were processed using the QIIME pipeline v 1.8.0 (Caporaso et al, 2010) with the settings described previously (Skvortsov et al, 2016). All sequences assigned to the non-bacterial "Unclassified" category and having similarity to the rRNA genes of chloroplasts were excluded from the subsequent analysis.

Identification of complete genomes

To detect end overlaps in the assembled contigs, the first 2,000bp of each contig were aligned against the whole contig's sequence. A contig was considered complete if a repeat of at least 150bp at its end was detected.

Analysis of contig dynamics

To assess the dynamics of individual viral genomes, the sequencing reads from each library were mapped onto sequences from the representative dataset using BBMap (percentage of nucleotide identity – 99, randomly selected best mapping site). The obtained number of reads mapped on a contig was normalised to the contig length and was additionally divided by the number of reads in a given library and multiplied by the mean value of the number of reads in 12 libraries. The resulting values were used as proxies of the relative abundances of viral genomes. To reduce the amount of information on the abundance of 8,950 contigs, peaks of abundance were determined. Relative abundances which were higher than the mean value of abundance for a particular genome were considered as belonging to a peak of abundance. A small fraction of the assembled contigs had two peaks of abundance - at the start of the period studied and at its end; these were considered as a single peak of abundance spanning the

148 winter-early spring period. To visualise the seasonal succession of viral genomes, peaks of
149 abundance were sorted and plotted using R (http://www.R-project.org/).
150 Analysis of the detectability of protein-based clusters in the environment by a metagenomics
151 method
152 To assess the number of reads in each sequencing library which could potentially belong to
153 the assembled contigs, all reads from each library were mapped onto contigs with 95%
154 nucleotide identity. A contig was considered undetectable in a given library if no reads
155 mapped onto it (coverage = 0.0). We then analysed the protein-based clusters (see below),
156 and considered a P-VC to be undetectable in a given library if all contigs comprising it were
157 undetectable in this library.
158 Clustering
159 For the clustering of viral contigs, a method developed by Lima-Mendez (Lima-Mendez et al,
160 2008) et al. was implemented. Briefly, the predicted protein sequences of contigs were
161 aligned against themselves ('all-to-all' protein blast search, threshold of 50 on bitscore) and
162 protein families were determined with the application of Markov cluster algorithm software
163 (MCL (Enright et al, 2002), inflation factor 1.2). Next, the pairwise comparison of shared
164 gene content between contigs was made using a hypergeometric formula, and significance
165 was calculated with correction for multiple comparisons (threshold of 0 on significance).
166 After that, the next round of clustering (MCL, inflation factor 1.1) generated groups of
167 related genomes. The inflation factor controls granularity of final clusters and as we analysed
168 community structure on two levels of similarity, for protein-based clusters (highest level of
169 organisation) we adjusted this parameter to maximise sizes of clusters. To obtain the clusters
170 of contigs sharing nucleotide homology, this method was adjusted and the protein blast
171 search was replaced by a nucleotide one. Thresholds were also adjusted and more strict
172 criteria were applied (a threshold value of 5 for significance and an inflation factor of 2 were

used for the second round of clustering). We then combined the results of these two clustering procedures in a single structure.

The third clustering was performed with the combined seeded sequences of isolated viruses (viral RefSeq, version 9/06/16), contigs assembled from the publicly available metagenomes and contigs assembled in this study with settings as for the first protein clustering. The clusters obtained, which included both types of contigs - long contigs of the Lough Neagh representative dataset, and seeded sequences - were transformed in pairs of long contigs and similar seeded genomes and assigned to the structure of the viral community generated in previous clustering procedures.

Assembly of publicly available freshwater metagenomes

Nine freshwater metagenomes were downloaded (Supplementary Table 1, Sheet 2). Metagenomes were assembled using IDBA-UD (kmer range 20-200, step - 10). Sequences longer than 10kb were combined and seeded to clustering.

Host-bacteriophage pairs prediction.

The software metaCRT (Rho et al, 2012) was used to predict CRISPR arrays in bacterial genomes (bacterial NCBI RefSeq, version of 22/08/2016). The sequences of spacers were collected, aligned against the set of long contigs, and only complete matches of the full length of spacers to contigs were allowed for the host prediction. Manual curation of predicted hosts was performed and links which included bacteria present among Lough Neagh OTU were left (Supplementary Table 3, Sheet 4).

Auxiliary metabolic gene identification

AMGs were considered to be genes that co-localised with ORFs of known viral origin on the same contig. To that end, contigs from all 12 libraries whose ontological annotation (Subsystems (Overbeek et al, 2005)) comprised the words “phage”, “terminase” or “capsid” were selected. Next, all functional annotations assigned to contigs selected in the previous

198 step were summarised. The category "Phages, Prophages, Transposable elements, Plasmids"
199 was removed from the final list of AMGs as it contains structural viral proteins and common
200 viral enzymes (Supplementary Table 4, Sheet 1).
201 To assess changes of gene content of reconstructed viral genomes in the environment
202 throughout the year, we evaluated and weighted the presence of functional categories of the
203 highest annotation level of SEED Subsystems for these genomes at each sample collection
204 time point. In order to do this, viral contigs were uploaded to MG-RAST server
205 (Supplementary Table 1, Sheet 1) for annotation, and annotations of the highest level were
206 collected for each contig. In each of the sample collection points, each functional annotation
207 was assigned a weight equal to the relative abundance of the contig that annotated feature
208 belonged to. Weights of all annotations of each particular functional category were summed,
209 normalised to the sum of all weights, and clustered with dist/hclust functions of R (Euclidean
210 distance, Ward clustering method.
211 Experimental verification of contigs
212 Experimental validation of the existence of DNA sequences of six contigs was performed
213 using PCR amplification of specific genome regions and subsequent partial resequencing of
214 amplicons from forward and reverse primers. The primers were designed with Primer-
215 BLAST (Ye et al, 2012) online software (Supplementary Table 5, Sheet 3 and Supplementary
216 Figure 1). For PCR amplification the same viral DNA samples were used as for the library
217 preparation for Illumina sequencing. The 25 µl of PCR mixture included 1 U of DreamTaq
218 DNA polymerase and its buffer (1x) (Thermo Fisher Scientific, Waltham, MA, USA), 0.2
219 mmol of each dNTP, 0.3 µmol of each primer and 8-10 ng of DNA template. PCR cycling
220 conditions were as follows: 1) initial denaturation at 95 $^{o}$C for 4 min, 2) denaturation at 95 $^{o}$C
221 for 30 s, 3) annealing at 60 $^{o}$C for 30 s, 3) elongation at 72 $^{o}$C for 7 min, 4) repeat steps 2-4
222 forty-five times, 5) final extension at 72 $^{o}$C for 4 min. The full volumes of PCR products

were loaded on 0.8% agarose gel. The lengths of amplicons were determined using the GeneRuler 1 kbp DNA ladder (Thermo Scientific) and products of required size were excised from the gel under UV light. DNA amplicons from agarose gels were extracted with High Pure PCR Product Purification kit (Roche Diagnostics, Rotkreuz, Switzerland) and sequenced at the University of Dundee DNA Sequencing and Services Facility.

Visual data exploration

To visualise pairwise genomic homology and similarity we used Easyfig v.2.2.2 (Sullivan et al, 2011).

The software package Gephi (Bastian et al, 2009) was used to visualise the results of the viral population clustering. To this end, the list of graphs (filtered pairwise comparisons of contigs with an estimation of their gene shared content) produced during DNA-based clustering was filtered in accordance to generated DNA-VCs (during this step all weak connections between contigs were removed). To the list obtained, graphs of protein clusters without DNA-VCs within them were added. These graphs were obtained from the protein-based clustering experiment. After that, a single random contig from each DNA-VCs within a given P-VCs was additionally connected to an artificial node as well as to all contigs unclustered into DNA-VCs within the same P-VC. All unique contigs, which remained fully unclustered, were transformed into a form of self-connected graphs and added to the final list of graphs which was loaded to Gephi. To generate the picture, the ForceAtlas2 algorithm was used.

Statistical analysis

Wilcoxon-Mann-Whitney test was used to compare highest abundances of two groups of contigs: with narrow form of peaks of abundances and with wide peaks ($U = 8269784.5$, $p < 0.01$). Spearman's rank correlation test was used to assess the strength and direction of correlations, with a value of $rho > 0.5$or $rho < -0.5$ considered as meaningful. Statistical

247 analysis was performed in R version 3.2.2 (http://www.Rproject.org/) and using Scipy (van
248 der Walt et al, 2011) packages for Python.

249 **Results and Discussion.**

250 **Succession of viral genotypes in Lough Neagh.**

251 To generate a representative dataset of viral genomic contigs that contains sequences of less
252 abundant viruses and viruses with pronounced seasonality, we applied a hybrid assembly
253 approach combining both assembly of individual metagenomic libraries and cross-assembly
254 (see Methods). The final dataset comprised 8,950 long contigs (≥ 7kb), which accounted for
255 59.2% of all reads. Among these contigs, 313 were considered to be complete genomes as
256 they had end overlaps (Supplementary Table 3, Sheet 2). The integrity of several assembled
257 contigs was verified experimentally using PCR amplification and partial resequencing by
258 Sanger's method (Supplementary Figure 1). These contigs were chosen mostly at random,
259 but included one complete small 7,148bp genome of a putative temperate phage (based on the
260 similarity of one its ORF with integrases), whose circular form was verified using PCR.
261 Another one was a contig encompassing a CRISPR array, the accurate assembly of which
262 was proved with resequencing.
263 To draw a picture of the annual succession of viruses, we determined the temporal dynamics
264 of all individual genomic contigs (Fig. 1, Supplementary Table 2, Sheet 1). For visual clarity
265 in Fig. 1, we have omitted some information and retained only data on abundances which
266 were higher than the mean value - peaks of abundance. Most viral contigs analysed (85.4%)
267 had a single peak of abundance during the year, and it was possible to distinguish two main
268 types - narrow (33% of all contigs) and wide (52%) peaks of abundance. Interestingly,
269 viruses with narrow peaks of abundance also were among the most abundant genomes in the
270 community (Wilcoxon-Mann-Whitney test, $p < 0.01$, see Methods). The detection of peaks of
271 the same genomes at the beginning and the end of the twelve-month period studied (Fig. 1)

suggests that this cycle of succession of viral species is annually repeated. After analysis of dynamic changes in contigs we assessed their presence in the environment during the year. This study demonstrated that only 39.1% of viruses (3,502 partial genomes) persisted in the lake throughout the year, while most viruses were undetectable by metagenomics methods at one or more time points. The characteristics of the dynamic changes in Lough Neagh viral populations should not be considered specific only to this particular environment; on the contrary, it is likely to be an instance of a universal phenomenon, reflecting processes common to different ecosystems on the global scale. For example, in a previous study of marine myoviruses it was demonstrated that during three consecutive years a number of viral genomes appeared only once a year at specific seasons and that only 25% of myoviruses persisted in the environment (Chow & Fuhrman, 2012). A study of viral dynamics in the hypersaline Lake Tyrell also revealed the presence of two types of viruses – those considered persistent and those detectable only at specific time points (Emerson et al, 2012; Emerson et al, 2013).

**Structure of the viral community**

Next we characterised the structure of the viral community. Reticulate classification of viral sequences allows estimation of the relatedness of genomes by assessing shared gene content (Lima-Mendez et al, 2008). This method uses comparisons of amino acid sequences, allowing grouping of viral genomes which do not necessarily have nucleotide homology (protein-based viral clusters, P-VCs) into clusters that approximately correspond to viral genera (Lima-Mendez et al, 2008; Roux et al, 2015; Roux et al, 2016; Paez-Espino et al, 2016). One of the goals of our analysis was to additionally divide assembled genomes within these clusters into subgroups of homologous sequences. To this end, we modified the method of reticulate classification and performed a second clustering using comparison of nucleotide sequences (DNA-based viral clusters, DNA-VCs). As the result of this, the majority of

297 contigs were organised into 732 P-VCs (Fig 2, Supplementary Table 3, Sheet 1) consisting of
298 2 to 696 members, while 1,198 contigs (13.4%) remained as singletons. Within the P-VCs,
299 sequences were arranged into sub clusters on the basis of sequence homology in DNA-VCs
300 (1,811 clusters in total, Supplementary Fig 2). The analysis of genome relatedness within this
301 double-clustered structure showed that the similarity of viral genomes within P-VCs varied,
302 which additionally characterises the community studied. For example, genomes within P-
303 VC_2 (Fig. 3) were very similar and retained some nucleotide homology across the whole
304 cluster/viral genus. This could point to the possibility that these viruses underwent gene
305 reshuffling more often than they accumulated point mutations. By contrast, genomes within
306 P-VC_20 (Fig. 3) are likely to have evolved under different constraints, as the genomes
307 detected were more distantly related even in smaller groups (DNA-VCs), retaining only
308 protein similarity between genomes from different DNA-based clusters.
309 The temporal dynamics of the clusters adds a new dimension to our understanding of viral
310 biodiversity. We explored how contigs, the majority of which had distinct seasonality, were
311 distributed between clusters and found that large P-VCs (with more than 20 partial genomes)
312 persisted during the year, although they could include DNA-VCs with specific seasonalities.
313 Thus, although certain genetic variants could appear for short periods only, the group of
314 viruses they belonged to could be detected throughout the whole year. At the same time,
315 smaller P-VCs could be abundant only during particular periods of the year (382 of all P-VCs
316 i.e. 52.2%; 51 of these included more than 4 contigs, Supplementary Table 3, Sheet 1) and
317 one-third of these (131 P-VCs) were undetectable by metagenomics technologies in other
318 periods (Methods). Moreover, we found that about one-fifth of all P-VCs (164, 22.4%) were
319 undetectable at specific time points.
320 **Identification of related sequences among known phages.**

To identify how the de novo assembled contigs were related to known viruses, complete viral genomes from RefSeq were seeded to a standard reticulate classification (see Methods). We also included in the analysis 488 long contigs (>10kb) assembled from nine published viral metagenomes originating from freshwater environments in different continents (Europe (Roux et al, 2012), North America (Green et al, 2015; Watkins et al, 2015), Africa (Fancello et al, 2013), Asia (Tseng et al, 2013), Supplementary Table 1, Sheet 2). The fact that only 48 RefSeq viruses were assigned to reconstructed viral genomes (Supplementary Table 3, Sheet 3) and 18 of them included in 19 DNA-VCs from our dataset reveals just how limited exploration of freshwater viral diversity has been. Among these were 8 species of Cellulophaga phages, 8 Pseudomonas phages and 7 cyanophages. We also identified one contig with similarity to an algal virus virophage - *Phaeocystis globosa* virus virophage (Supplementary Table 3, Sheet 3). The seeding of long contigs assembled from other freshwater metagenomes allowed us to determine that 106 of them (21.7%) were related to the viruses in Lough Neagh (Supplementary Table 3, Sheet 1). In total, 69 DNA-VCs (from 40 P-VCs) recruited contigs from other freshwater environments. One of the P-VCs (P-VC_19) seemed to represent a "core freshwater cluster" of genomes, as it recruited viral sequences from five freshwater reservoirs from very distant sites: Lough Neagh (British Isles), Lake Michigan (North America), Lakes Pavin and Bourget (Continental Europe), and the Feitsui freshwater reservoir in Taiwan (Asia). Several sequences from this "core" cluster were related to *Cellulophaga* phage 46:1 (Holmfeldt et al, 2013) (Supplementary Table 3, Sheet 1).

Although, the method of co-clustering of viral genomes allows to detect more distant relatives among known sequences, we additionally explored results of reads-mapping approach of MetaVir pipeline to identify eukaryotic viruses which were less likely to assemble due to predominance of bacteriophages in the environment. The highest number of

reads of eukaryotic viruses were assigned to the *Phycodnaviridae* family of algae viruses represented by all genera with *Chlorovirus* as the most abundant one. Among other viruses of eukaryotic organisms, sequences for several gigantic viruses of family *Mimiviridae*, such as amoebic *Acanthamoeba polyphaga* moumouvirus and flagellate *Cafeteria roenbergensis* virus, were found. Sequences related to viruses of vertebrate and invertebrate animals of families *Iridoviridae*, *Herpesvirales* and *Poxviridae* were detected as well.

**Dynamic relationships of viruses and their predicted hosts.**

To gain insights into the biology of the reconstructed viruses, we predicted their bacterial hosts using a sequence-based bioinformatics method of CRISPR matching. In a recent benchmarking analysis CRISPR matches yielded the highest accuracy (92%) of all tested bioinformatics approaches designed to link phages to their hosts (Edwards et al, 2016). Throughout the year we generated structural profiles of the bacterial community using methods of amplicon-based metagenomics (Methods, Supplementary Figure 3, Supplementary Table 2, Sheet 2). Among known caveats of this approach is that the resolution provided by 16S amplicons is not necessarily sufficient to distinguish ecotypes, which have identical 16S sequences, but different genomes and may demonstrate individually distinct dynamics in the ecosystem. To link reconstructed phage genomes to their potential hosts, we identified CRISPR arrays in the complete genome sequences of bacterial species that were closely related to the OTUs detected in Lough Neagh by using selected bacterial genomes from the database, and matched those spacers to our reconstructed viral genomes (see Methods). Hosts were predicted for 225 of the 8,950 reconstructed viral genomes. For several contigs up to three potential bacterial hosts of different OTUs were assigned (possible viral generalists), therefore in total we found 260 phage-host pairs (Supplementary Table 3, Sheet 4). Although the database bacteria were isolated from different locations and have never been exposed to the Lough Neagh phages, we presumed that their recent ancestors

were indeed infected by close relatives of these phages, as evidenced by the 100% identical CRISPR spacers. Viruses tend to be species or strain-specific, and when they do change their host tropism, they mostly switch to taxonomically very closely related hosts (Popa et al, 2017). This is an indirect approach to predict phage-host pairs, but we believe that it provides accurate insights into phage-bacteria relationships for the minority of cases where hits were found.

The contigs with the hosts assigned belonged to 131 DNA-VCs of 97 P-VCs (13.4%). The analysis of P-VCs showed that, although viruses from a given cluster usually infect a single dominant bacterial taxonomic group, there were also clusters with predicted hosts from up to five different classes and two different phyla. This finding supports the idea that, although the majority of genetically related viruses have a narrow host repertoire, there are also generalist viruses and viral genera, which can prey on hosts across bacterial taxonomic borders (Malki et al, 2015, Knowles et al, 2016; Peters et al, 2015, Roux et al, 2016).

We studied dynamic changes in viral contigs and presumed hosts (OTUs) to identify possible patterns of their interactions in a natural environment. In order to do that, we plotted the distribution of highest abundance of reconstructed viral genomes in relation to the maximum of corresponding bacterial abundance (Fig. 4). In accordance with the 'Kill-the-Winner' (KtW) model of host-pathogen relationships (Thingstad, 2000), the dynamics of bacteria and their viruses are co-dependent, and the peak of abundance of a virus should appear with some delay after the peak of abundance of its host. The correlational analysis (Spearman's rank correlation, rho > 0.5) of relationships of identified viral contig – bacterial OTU pairs demonstrated that 54 pairs behaved in accordance with the KtW model (20.8%, see Methods) and the dynamic changes of 28 other phage-host pairs coincided (31.5% in total). But this plot demonstrates that many viruses peaked before their hosts. We performed correlational analysis and found that in 43 pairs (16.5%) the increase of viral abundance was indeed

396 followed by the increase of host density. To our knowledge, this is one of the few examples
397 when viral abundance peaks occurring before the peaks of their cognate host have been
398 observed in natural environment.
399 Next, we performed an investigation of the existing literature, looking for evidence where
400 this counter-intuitive pattern may have been registered. Wilson et al. presented time series
401 data of marine mesocosms where, after addition of phosphorus to the environment, and
402 before the development of a peak of cyanobacterial abundance as a response, there was a
403 distinguishable high peak of abundance of viral particles (Wilson et al, 1998). Similarly, in a
404 time-series study of marine Synechococcus and cyanophage populations, although this
405 observation was outside the scope of the paper, preceding peaks of viral abundances were
406 noticeable and were repeated on several occasions over the period studied (McDaniel et al,
407 2002).
408 Moreover, this dynamic pattern was modelled for situation of effective defence of prey from
409 low-offence predators (Cortez & Weitz, 2014). We offer several possible mechanisms of
410 such defence that might explain the observed dynamics. First, they might be explained by the
411 development of resistance of bacteria to the phage, for example by acquisition of CRISPR
412 spacers or modification of their receptor binding proteins, facilitating subsequent expansion
413 of the bacterial population. However, mechanisms of resistance acquisition can also be due to
414 super-infection exclusion caused by the switch of phages from the lytic to the temperate state.
415 Recently, Knowles et al. proposed an extension to the KtW model – the Piggyback-the-
416 Winner (PtW) model, in accordance with which 'temperateness is favoured at high host
417 densities as viruses exploit their hosts through lysogeny rather than killing them' (Knowles et
418 al, 2016). We might expect that the dynamic pattern identified could be a result of phage-host
419 interactions in accordance with this PtW model. Moreover, it was recently revealed that some
420 viruses can communicate with each other via short quorum-sensing peptides, where an

increase of the peptide concentration causes switch of temperate phages from the lytic to the lysogenic state (Erez et al, 2017). It is possible that this mechanism could also explain the "early" loss of viruses from the environment, as observed in our study.

**Environmental parameters in Lough Neagh.**

To discover as many drivers of viral community changes as possible, we characterised bacterial community composition and environmental parameters in the lake's ecosystem (Supplementary Table 2, Sheet 2 and Supplementary Table 5, Sheet 1). Predominance of cyanobacteria in eutrophic Lough Neagh was detected in summertime (Supplementary Fig. 3). Comparative analysis of the dynamics of the bacterial populations and changes in physical and chemical parameters showed that temperature was likely to be the main driver of changes in the bacterial community under study (Spearman's rank correlation, rho > 0.5, Supplementary Table 5, Sheet 1). We also found that, surprisingly, the bacterial community did not react to changes in phosphorus concentration - the main limiting factor for growth of microbial populations in freshwater environments (Doering et al, 1995; Correll, 1999). Apparently, in this eutrophic lacustrine ecosystem the main limiting factor is different, which is in accordance with previous findings that in Lough Neagh nitrogen loading can have a stronger long-term impact than phosphorus on lake eutrophication (Buntig et al, 2007).

**Viral auxiliary metabolic genes and their changes throughout the year.**

Viruses can carry auxiliary metabolic genes (AMGs) that augment their fitness by affecting host metabolism (Breitbart et al, 2007). As it is not possible to exclude the occurrence of bacterial genes caused by generalised transduction events and the presence of gene transfer agents in phage metagenomes, we applied strict criteria for the detection of phage-associated AMGs. Metabolic genes were considered as AMGs only if they were co-localised on the same contig with open reading frames (ORFs) having similarity to known phage genes (such as structural genes, see Methods). Contigs from all 12 assembled libraries were analysed and

189 phage-associated AMGs were identified (Supplementary Fig. 4, Supplementary Table 4, Sheet 1). The attribution of AMGs to SEED subsystems showed that freshwater viruses in Lough Neagh had acquired genes from a wide variety of metabolic pathways, related to almost all aspects of bacterial life, since genes from 25 out of 30 of subsystems were found in viral genomes.

To assess how the appearance of various genes (functional categories of SEED subsystems (Overbeek et al, 2005)) in the viral population depended on the dynamics of the bacterial community and on environmental parameters, a correlational analysis was performed (Spearman's rank correlation, rho > 0.5, Supplementary Table 4, Sheet 2). An increase in relative abundance of "Genes of temperate phages" in the summer viral community was detected, supporting findings from previous studies obtained by using different methods (Knowles et al, 2016; Laybourn-Parry et al, 2007; Palesse et al, 2014). We also identified correlations between the appearance of genes of "Oxidative stress response regulation" in the summertime and Cyanobacteria changes and alkalinity fluctuations (Fig 5B). Cyanobacteria undergo oxidative stress more often than heterotrophic bacteria due to their photosynthetic ability (Latifi et al, 2009), and it was shown that marine cyanophages can carry genes involved in photoprotection, such as those encoding high light inducible proteins (Ma et al, 2014). "Oxidative stress response regulation" genes of freshwater viruses identified in this study included a wide range of molecules guarding living organisms from oxidative damage: iron and manganese superoxide dismutases, peroxidase, catalase, ferroxidase, rubrerythrin etc. In contrast with high light inducible proteins of marine cyanophages, which are tightly connected with photosystem formation and functioning (Komenda & Sobotka, 2016), antioxidant defence genes of freshwaters are more general and include cytoplasmic, mitochondrial and chloroplast-associated molecules. Other notable correlations were detected between the abundance of the Bacteroidetes phylum, the Verrucomicrobiae class and genes

of "Quorum sensing and biofilm formation" (Fig 5C). The relative abundance of this functional category did not correlate with any other taxa or any environmental parameter, suggesting that this type of phage manipulation is specific to these clades.

To further investigate the seasonal dependence of viral functional potential we annotated separately reconstructed viral genomes on MG-RAST server and clustered functional annotations, weighted by contig relative abundance in the community (see Methods, Fig. 5A). We identified that reconstructed viral genomes clearly differed in the winter-early spring and summer-autumn periods by functions, these two groups being largely subdivided in accordance with calendar seasons. These findings could additionally point to the specialisation of viruses to their hosts through acquisition of specific AMGs.

**Conclusions.**

Overall, this study changes our understanding of viral diversity by demonstrating the transient nature of most viral groups of genomes in an ecosystem. This variation of the whole metagenomic content of the environment between different seasons/months should also be considered when assessing the criteria for the sampling completeness of an ecosystem. Visualisation of the genetic relationships between viruses further characterises the community as a whole and points to the diversity of evolutionary constraints in a natural environment. Besides providing much-needed insight into freshwater viral sequence diversity and ecosystem organisation, our research offers a basis for long-term studies on the stability of individual viral genomes, on the repeatability of seasonal cycles, and on their interplay with bacterial host communities. In our study, we analysed only enveloped DNA viruses existing as viroid particles in the environment. Previous studies have highlighted that viruses can also subsist inside their host cell for prolonged periods of time, so it will be interesting to analyse time series of combined free-viroid and induced viromes side-by-side (Maurice et al,

2011). Moreover, including time-series experiments of RNA viruses can also provide complementary insight into the dynamics of viral communities in the future.

**Acknowledgements**

We thank Dr Catherine Watson and Ms Hanna Cromie for her help with the collection of the samples. The project was funded by Leverhulme Trust Grant RPG-2013-040. B.E.D. and K.A. were supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004. All data will be available from SRA, MetaVir and MG-RAST databases immediately after the publication date (for accession numbers see Supplementary Table 1, Sheet 1).

**Conflict of Interest**

The authors declare no conflicts of interest.

Supplementary information is available at The ISME Journal's website

508

509 **References.**


510 1. Bastian M, Heymann S, Jacomy M (2009). Gephi: an open source software for
511 exploring and manipulating networks. International AAAI Conference on
512 Weblogs and Social Media.
513 2. Bolduc B, Wirth JF, Mazurie A, Young MJ (2015). Viral assemblage composition
514 in Yellowstone acidic hot springs assessed by network analysis. ISME J 9:2162-
515 2177
516 3. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. (2007). Exploring the vast
517 diversity of marine viruses. Oceanography 20:135-139.
518 4. Buntig L, Leavitt PR, Gibson CE, McGee EJ, Hall VA. (2007). Degradation of
519 water quality in Lough Neagh, Northern Ireland, by diffuse nitrogen flux from a
520 phosphorus-rich catchment. Limnol Oceanogr 52:354–369.
521 5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK
522 et al. (2010). QIIME allows analysis of high-throughput community sequencing
523 data. Nature Methods 7:335-336.
524 6. Chow CE, Fuhrman JA. (2012). Seasonality and monthly dynamics of marine
525 myovirus communities. Environ Microbiol 14:2171-2183
526 7. Correll DL. (1999). Phosphorus: a rate limiting nutrient in surface waters. Poult
527 Sci 78:674-682.
528 8. Cortez MH, Weitz JS. (2014). Coevolution can reverse predator-prey cycles.
529 PNAS 111:7486-7491.
530 9. Doering PH, Oviatt CA, Nowicki BL, Klos EG, Reed LW. (1995). Phosphorus
531 and nitrogen limitation of primary production in a stimulated estuarine gradient.
532 Mar Ecol Prog Ser 124:271-287.

10. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE (2016). Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev 40:258-272

11. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbial 78:6309-6320

12. Emerson JB, Thomas BC, Andrade K, Heidelberg KB, Banfield JF (2013). New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. Appl Environ Microbial 79:6755-6764.

13. Enright AJ, Van Dongen S, Ouzounis CA. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575-1584.

14. Erez Z, Steinberger-Levy I, Shamir M, Doron S, Stokar-Avihail A, Peleg Y et al. (2017). Communication between viruses guides lysis-lysogeny decisions. Nature 541:488-493.

15. Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D et al. (2013). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. ISME J 7:359-369

16. Green JC, Rahman F, Saxton MA, Williamson KE (2015). Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA Aquat Microb Ecol 75:117-128

17. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, Verberkmoes NC et al. (2013). Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci USA 110:12798-12803.

18. Huang X, Madan A. (1999). CAP3: A DNA sequence assembly program. Genome Res. 9, 868-877 (1999).

19. King AMQ, Lefkowitz E, Adams MJ, Carstens EB (2011) Virus taxonomy: ninth report of the international committee on taxonomy of viruses. Elsevier:Amsterdam.

20. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG et al. (2016). Lytic to temperate switching of viral communities. Nature 531:466-470.

21. Komenda J, Sobotka R, (2016). Cyanobacterial high-light-inducible proteins – protectors of chlorophyll-protein synthesis and assembly. BBA-Bioenergetics 1857:288-295.

22. Latifi A, Ruiz M, Zhang CC. (2009). Oxidative stress in cyanobacteria. FEMS Microbiol Rev 33:258-278.

23. Laybourn-Parry J, Marshall WA, Madan NJ. (2007). Viral dynamics and patterns of lysogeny in saline Antarctic lakes. Polar Biol 30:351-358.

24. Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006).

25. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 25:762–777.

26. Ma Y, Allen LZ, Palenik B. (2014). Diversity and genome dynamics of marine cyanophages using metagenomic analyses. Environ Microbiol Rep 6:583-594.

27. Malki K, Kula A, Bruder K, Sible E, Hatzopoulos T, Steidel S, Watkins SC, Putonti C. (2015) Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. Virology J 12:164

582 28. Maurice CF, Mouillot D, Bettarel Y, Wit RD, Sarmento H, Bouvier T. (2011).
583 Disentangling the relative influence of bacterioplankton phylogeny and
584 metabolism on lysogeny in reservoirs and lagoons. ISME J 5:831-843.
585 29. McDaniel L, Houchin LA, Williamson SJ, Paul JH (2002). Lysogeny in marine
586 Synechococcus. Nature 415, 496 (2002)
587 30. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M et al. (2008).
588 The metagenomics RAST server – a public resource for the automatic
589 phylogenetic and functional analysis of metagenomes. BMC Bioinformatics
590 9:386.
591 31. Noguchi H, Taniguchi T, Itoh T. (2008). MetaGeneAnnotator: detecting species-
592 specific patterns of ribosomal binding site for precise gene prediction in
593 anonymous prokaryotic and phage genomes. DNA Res 15:387-396.
594 32. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M et al.
595 (2005). The subsystems approach to genome annotation and its use in the project
596 to annotate 1000 genomes. Nucleic Acids Res 33:5691-5702.
597 33. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M,
598 Mikhailova N et al. (2016). Uncovering Earth's virome. Nature 536:425–430.
599 34. Palesse S, Colombet J, Pradeep Ram AS, Sime-Ngando T (2014). Linking host
600 prokaryotic physiology to viral lifestyle dynamics in a temperate freshwater lake
601 (Lake Pavin, France). Microb Ecol 68:740-750.
602 35. Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo
603 assembler for single-cell and metagenomic sequencing data with highly uneven
604 depth. Bioinformatics 28:1420-1428.

36. Peters DL, Lynch KH, Stothard P, Dennis JJ. (2015). The isolation and characterisation of two Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. BMC Genomics 16:664.

37. Rho M, Wu Y, Tang H, Doak TG, Ye Y. (2012). Diverse CRISPRs evolving in human microbiomes. PLoS Genet 8: e1002441.

38. Popa O, Landan G, Dagan T. (2017). Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J 11:543-554.

39. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A et al, (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689-693

40. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS ONE 7:e33641

41. Roux S, Hallam SJ, Woyke T, Sullivan MB. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife 4:1–20 (2015).

42. Roux S, Tournayre J, Mahul JA, Debroas D, Enault F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15:76.

43. Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, Watson C, Arkhipova K, Lavigne R, Kulakov LA. (2016). Metagenomic characterisation of the viral community of Lough Neagh, the largest freshwater lake in Ireland. PLoS ONE 11:e0150361.

44. Sullivan MJ, Petty NK, Beatson SA. (2011). Easyfig: a genome comparison visualizer. Bioinformatics 27:1009-1010.

630 45. Thingstad TF. (2000). Elements of a theory for the mechanisms controlling
631 abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic
632 systems. Limnol Oceanogr 45:1320-1328
633 46. Tseng CH, Chiang PW, Shiah FK, Chen YL, Liou JR, Hsu TC et al. (2013).
634 Microbial and viral metagenomes of a subtropical freshwater reservoir subject to
635 climatic disturbances. ISME J 7:2374-2386
636 47. van der Walt S, Colbert SC, Varoquaux G. (2011). The NumPy array: a structure
637 for efficient numerical computation. Compu Sci Eng 13:22-30.
638 48. Wang MN, Ge XY, Wu YQ, Yang XL, Tan B, Zhang YJ et al. (2015). Genetic
639 diversity and temporal dynamics of phytoplankton viruses in East Lake. China
640 Virol Sin 30:290-300
641 49. Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J et al. (2015).
642 Assessment of a metaviromic dataset generated from nearshore Lake Michigan.
643 Mar Freshwater Res 67:1700-1708.
644 50. Wilson WH, Turner S, Mann NH. (1998). Population dynamics of phytoplankton
645 and viruses in a phosphate-limited mesocosm and their effect on DMSP and DMS
646 production. Estuar Coast Shelf Sci 46:49-59.
647 51. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. (2012).
648 Primer-BLAST: A tool to design target-specific primers for polymerase chain
649 reaction. BMC Bioinformatics. 13:134.
650 52. Yeo BH, Gin KY. (2015). Population dynamics of cyanomyovirus in a tropical
651 eutrophic reservoir. Microbes Environ 30:12-20
652

**Figure legends.**

**Figure 1. Succession of 8,950 assembled contigs throughout a year.** Each row in the left panel of the picture presents information about peaks of abundance for individual contigs. For each library, a dot was placed if the peak of abundance for a particular contig was attributed to this library; otherwise, an empty space was left. Jitter was applied to distribute dots belonging to different contigs within a single column. The right panel schematically depicts dynamic changes in individual contigs to provide illustrations of different observed cases. Contig identification numbers are specific to this Figure and do not correspond to contig IDs used elsewhere in the study.

**Figure 2. A.** An overview of the viral sequence space organisation in the community. Each dot represents an individual from 8,950 assembled contigs. These contigs can be i) genetically unique and fully unclustered, ii) clustered into DNA-VCs (middle-size clustering level), iii) clustered into P-VC being within DNA-VCs or being unclustered within DNA-VCs (large-size clustering level).

The outer ring of light grey dots is constituted by unique individual genomes that are not members of P-VCs (variant i). Each separate group of dots within the inner circle represents an individual P-VC (variant iii). All DNA-VCs as well as all unclustered contigs within each P-VC were joined to an artificial central node. To avoid confusion with colours, larger P-VCs were arbitrary coloured to provide more information about their inner structure. Contigs comprising DNA-VCs within P-VCs are coloured either in orange (relatively bigger) or in dark grey (relatively smaller), while unclustered into DNA-VCs contigs are coloured in green. By default, a dark grey colour is used for all contigs within other P-VCs.

Three distinct types of clusters are indicated: **B.** P-VC which includes sequences with nucleotide homology organised in DNA-VCs, as well as unclustered genomes (mixed type);

678 **C.** P-VC which aggregates (mostly) unclustered in DNA-VCs genomes (contigs within these
679 clusters have only similarity at the protein level); **D**. P-VC which comprises mostly DNA-
680 VCs.
681 **Figure 3. Examples of genome relatedness within the double-clustered structure of the**
682 **viral sequence space organisation.** Full-length nucleotide (a,b) and amino-acid (c)
683 alignments of representatives from DNA-VCs of two P-VCs are shown on the left side of the
684 Figure, the degree of homology between aligned fragments is colour-coded. All genomes
685 presented are complete, if not stated otherwise. The dynamic changes in abundance of these
686 individual genomes are shown on the right. a. Genomes preserving nucleotide homology
687 between different DNA-VCs within a single P-VC. Two upper genomes were clustered into
688 one DNA-VC, while the other 9 genomes belonged to different DNA-VCs within one P-
689 VC_2. b. Genome alignments of three members of DNA-VC-20. c. Genomic map of protein
690 similarity (tblastx) between three representatives from different DNA-VCs (including DNA-
691 VC_20 from panel b) within a single P-VC_20. These genomic sequences do not have
692 nucleotide homology.
693 **Figure 4. Positions of the peaks of abundance for 260 viral genomes in relation to**
694 **maximum abundance of their predicted host bacteria.** The numbers of viruses in groups,
695 organised by the distance (in sample collection intervals) of their peaks of abundance from
696 the peaks of abundance of host bacteria, are plotted as black dots (•), while host abundance
697 maximum is used as a reference point and represented by a dashed line (---).
698
699 **Figure 5. A.** Clustering of 12 viromes based on functional annotations of 8,950 reconstructed
700 genomes, weighted by genome's abundance. **B.** Dynamic changes of relative abundance of
701 Cyanobacteria (black, rho = 0.6) and “Regulation of oxidative stress response” functional
702 category genes (red). **C.** Dynamic changes of relative abundance of the Bacteroidetes phylum

703 (blue, rho = 0.68), the Verrucomicrobiae class (black, rho = 0.67) and genes of "Quorum
704 sensing and biofilm formation" (red) functional category
705
706 **Author contributions**
707 L.A.K., T.S. and K.A. designed the study, T.S. processed water samples, K.A. and T.S.
708 performed bioinformatics analyses of data, K.A., T.S., J.P.Q., J.W.G., C.C.R.A., B.E.D.,
709 Y.M., C.W. and L.A.K. discussed results and wrote and edited the manuscript.
710

**Legends for supplementary files.**

**Supplementary Figure 1.** Genomic maps of the verified experimentally assembled contigs. In the middle part of genomic maps, the fragments that were amplified and resequenced from forward and reverse primers and their overlaps are marked in pink and burgundy, respectively. Genes above the middle part are located on "plus" chain, below – on the "minus" chain of DNA.

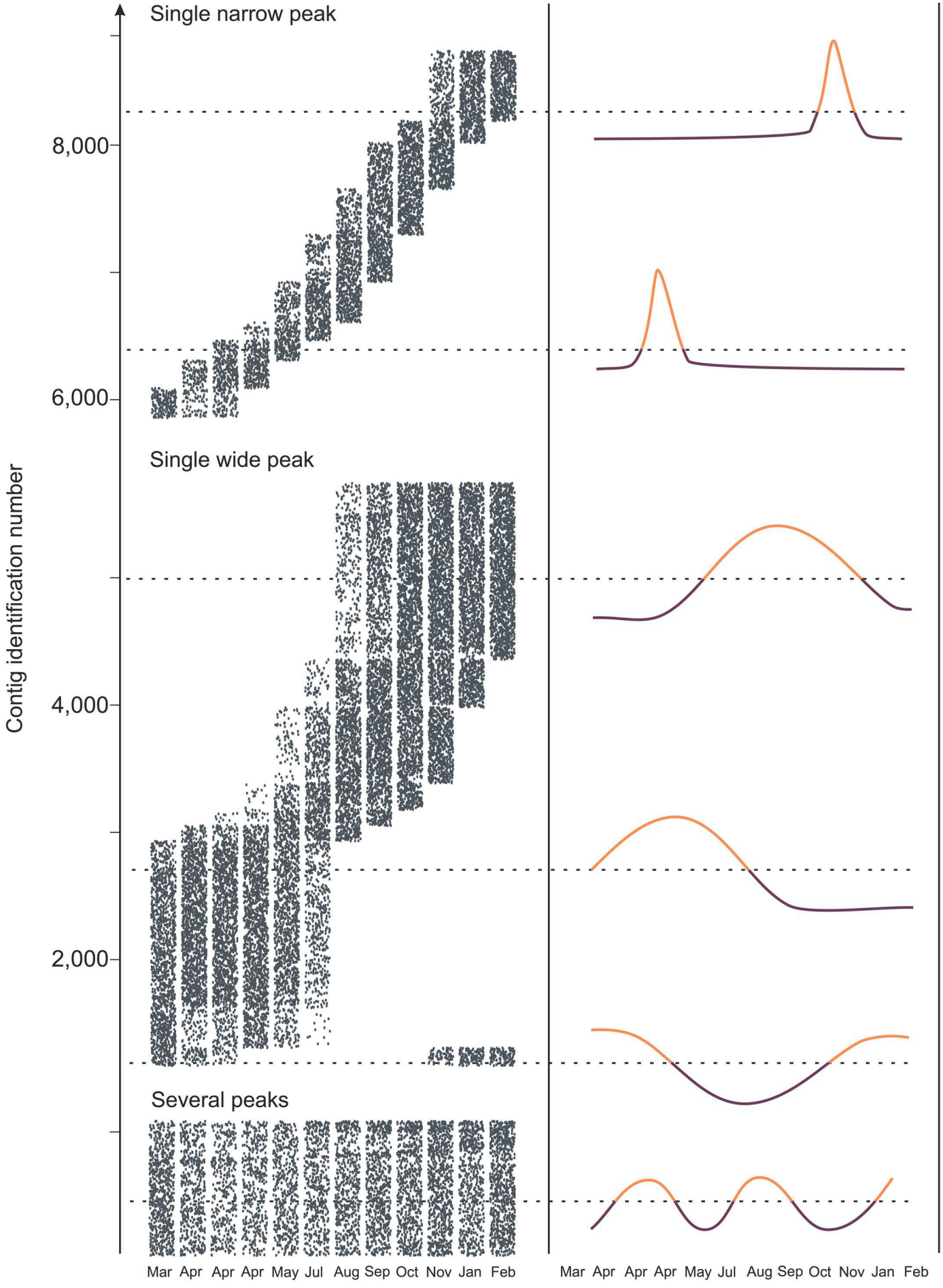
Supplementary **Figure 2. Quantitative characteristics of generated protein-based clusters. A.** Distribution of numbers of DNA-VC per P-VC. **B**. Distribution of numbers of unclustered contigs per P-VC.

**Supplementary Figure 3.** Dynamic changes in six major bacteria phyla. Relative abundances of the six most abundant bacteria phyla in the water column of Lough Neagh are presented. This analysis included an additional sample collected in November 2013.

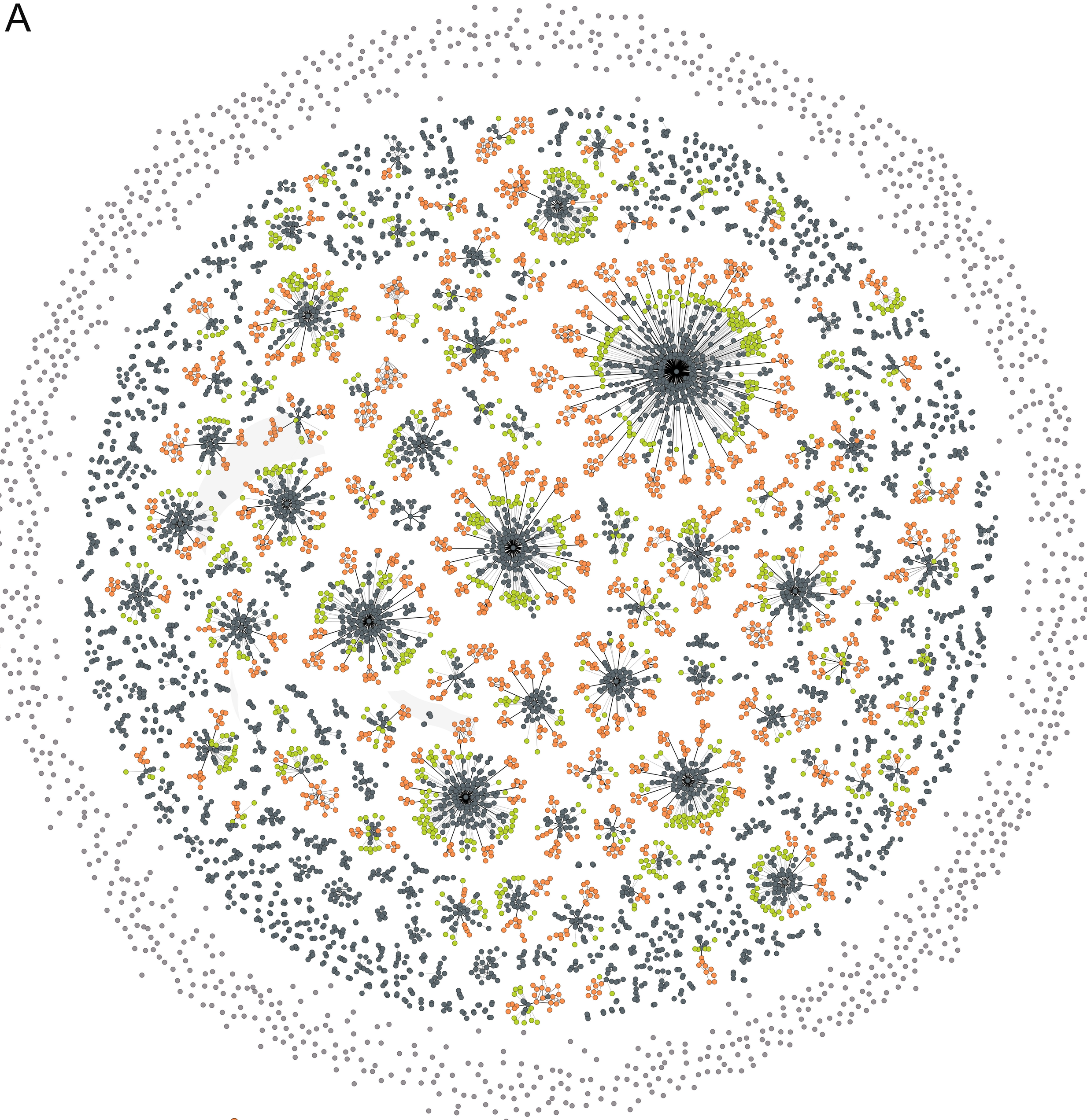
**Supplementary Figure 4.** Functional profiling of assembled viral libraries and AMG. The distribution of functional categories across 12 assembled libraries is shown on the left. For each category, the mean relative abundance and standard deviation is given. The "Phages, Prophages, Transposable elements, Plasmids" category (60.3 ± 8.2% of reads) is excluded from the bar chart. The distribution of AMGs (in this study: functional genes that are co-localised with phage genes) by functional categories is shown on the right. Proportions of given functional categories are presented. ND – not detected.

**Supplementary Table 1. Sheet 1.** Metagenomic data sets generated in this study and data availability. Raw reads and assembled contigs were deposited to SRA, MetaVir and MG-RAST for archival storage and/or analysis. **Sheet 2.** List of used freshwater metagenomes. This file includes accession numbers, references, number of assembled long contigs from

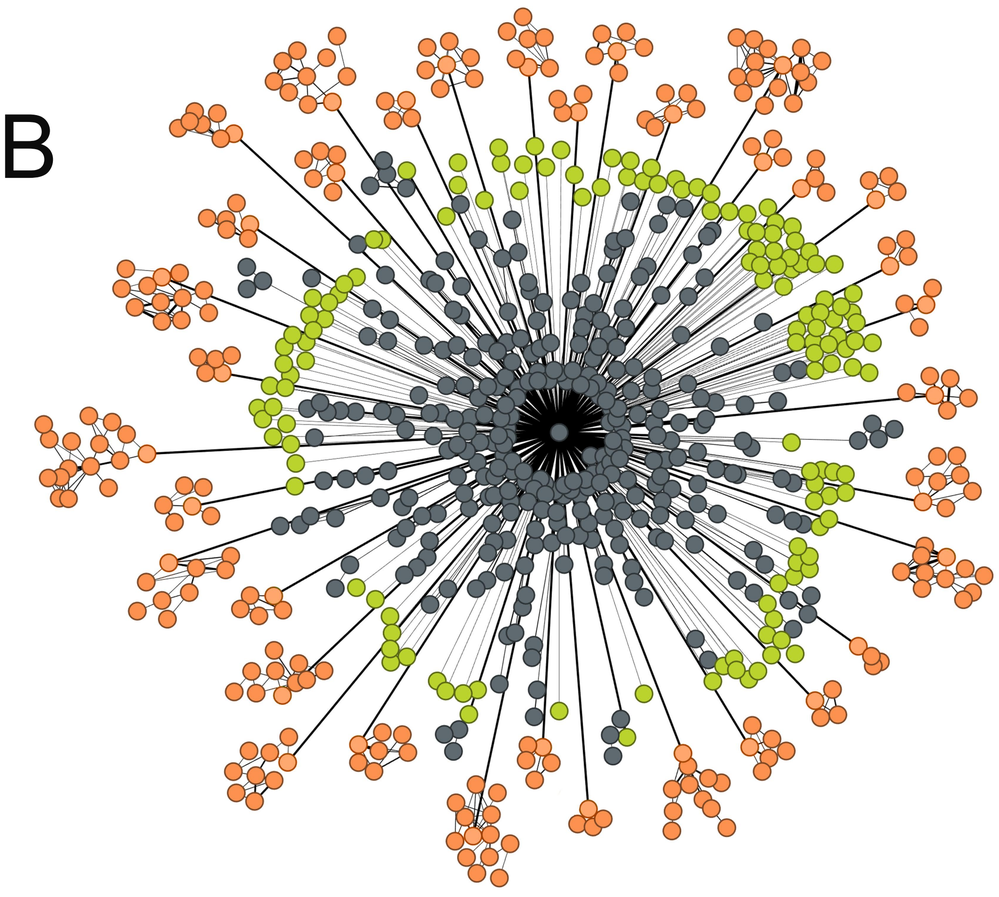
735 publicly available freshwater metagenomes and number of contigs resembling long genomes
736 from Lough Neagh.
737 **Supplementary Table 2. Sheet 1.** Relative abundance of assembled long contigs. **Sheet 2**.
738 Relative abundances of bacterial OTUs. **Sheet 3.** Percentage of contig length covered by
739 reads in each library.
740 **Supplementary Table 3. Sheet 1.** Summary statistics of viral clusters. For each P-VC,
741 DNA-VCs are listed and the number of contigs forming the clusters is given. The results of
742 co-clustering with viral genomes from NCBI RefSeq and contigs assembled from other
743 freshwater metagenomes are also included. The information about the presence of complete
744 viral genomes within clusters is provided where available. **Sheet 2.** Circular genomes
745 summary. This table contains information about the length of circular contigs and their
746 similarity to genomes of known bacteriophages. **Sheet 3.** This table contains information
747 about which assembled contigs co-clustered with which known viruses from RefSeq
748 database. **Sheet 4**. List of predicted host-bacteriophage links.
749 **Supplementary Table 4. Sheet 1.** List of identified AMGs. AMGs were grouped in
750 accordance with SEED Subsystems ontological classification and four levels of grouping are
751 indicated. **Sheet 2.** Correlations of appearance of genes with specific functions in the viral
752 population with changes in environmental parameters and bacterial abundance (Spearman'
753 rank correlations, rho > 0.5).
754 **Supplementary Table 5. Sheet 1.** Correlations (Spearman' rank correlations, rho > 0.5 and
755 rho < -0.5) between changes of bacterial relative abundance and environmental parameters.
756 **Sheet 2.** Physicochemical parameters of the lake at the times of sample collection. **Sheet 3.**
757 Sequences of primers used for experimental verification of assembled contigs.
758
759


Single narrow peak
Single wide peak
Several peaks
Contig identification number
8,000
6,000
4,000
2,000
Mar Apr Apr Apr May Jul Aug Sep Oct Nov Jan Feb
Mar Apr Apr Apr May Jul Aug Sep Oct Nov Jan Feb

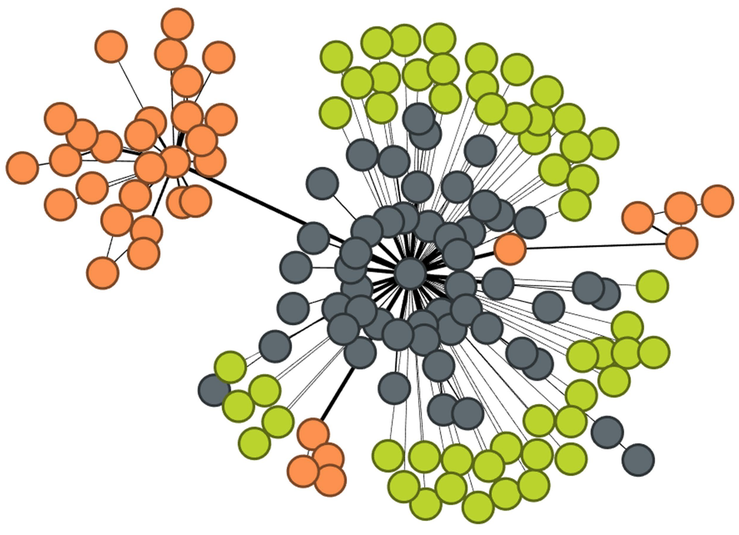
A



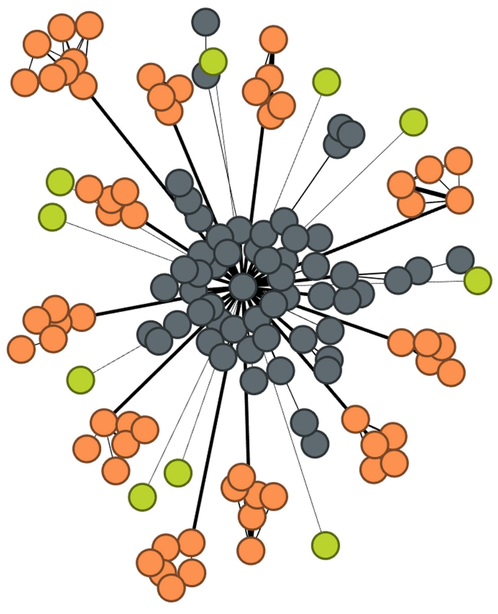
B



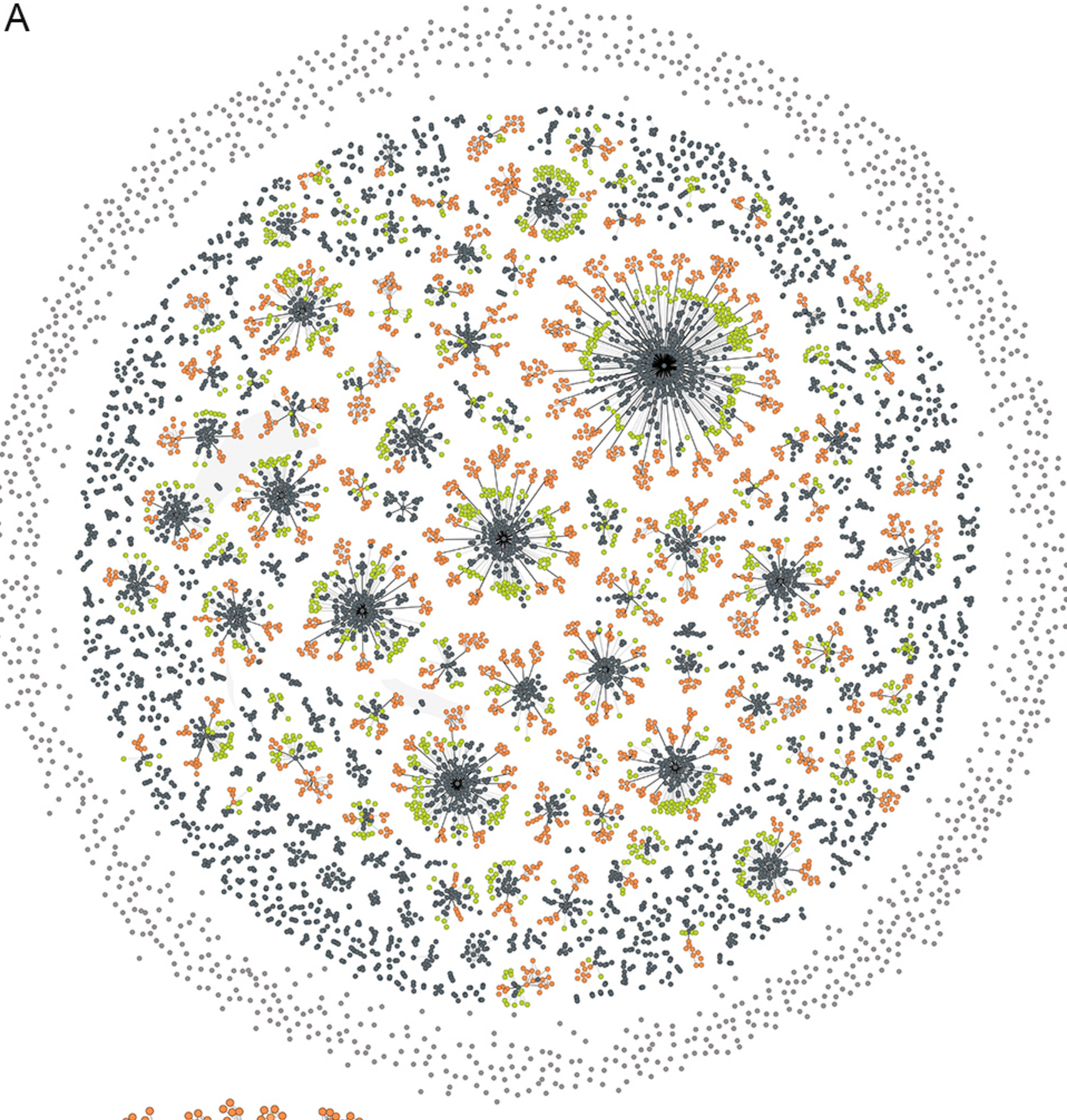
C



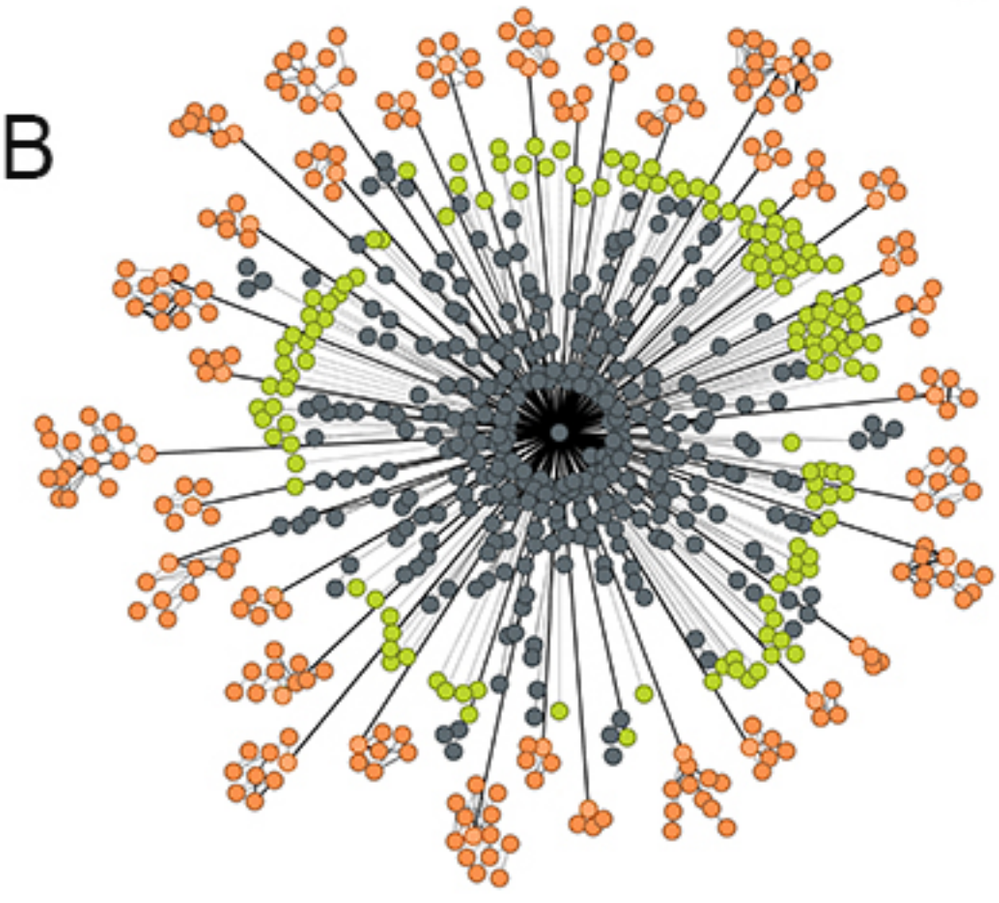
D



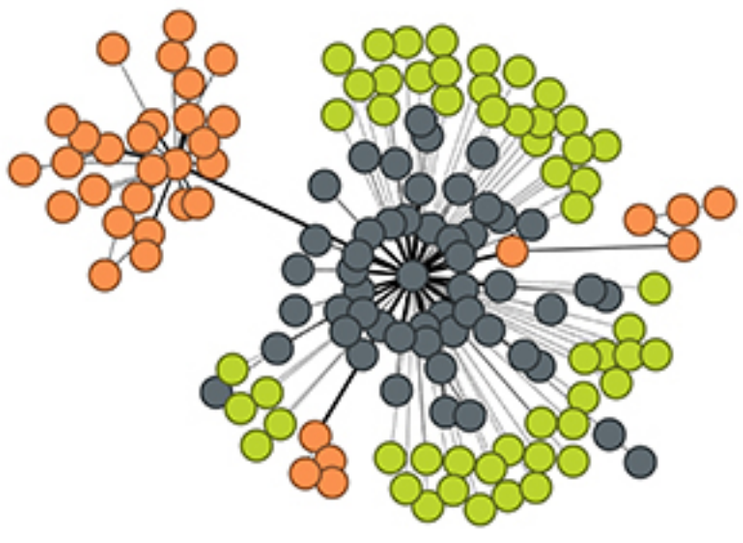
A



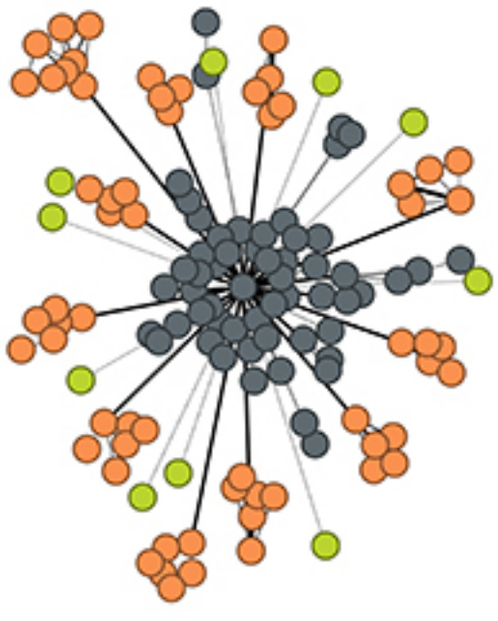
B

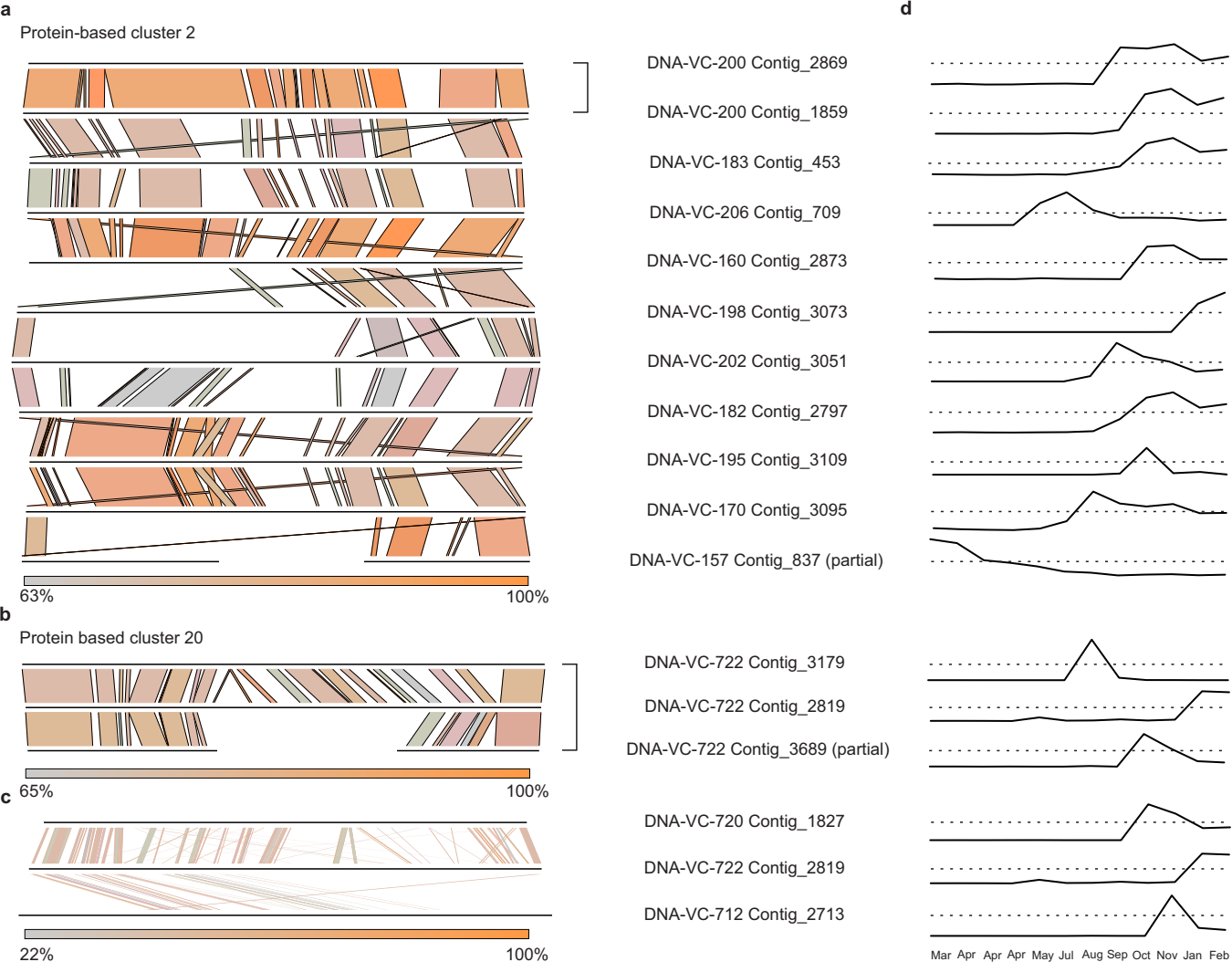


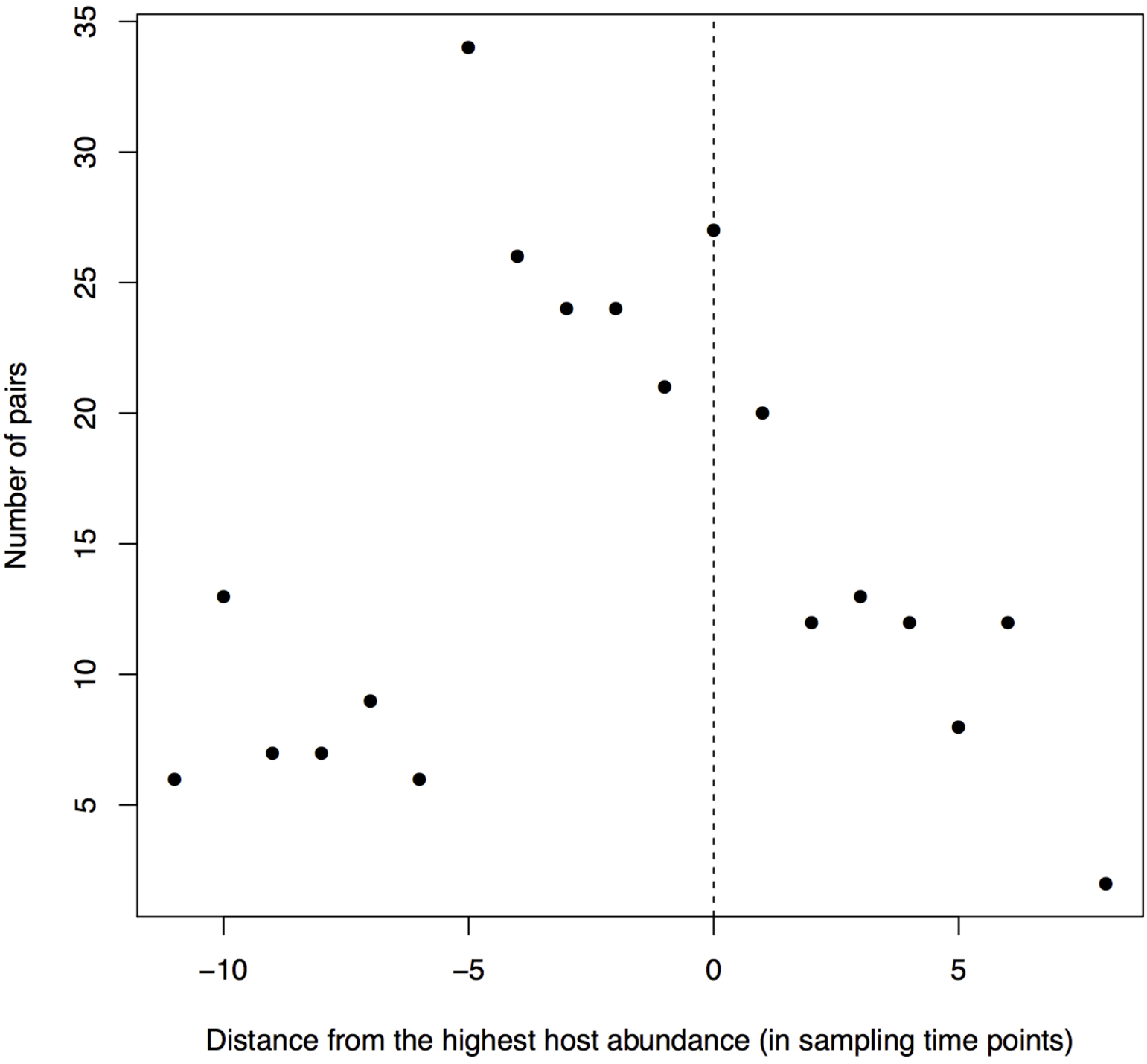
C

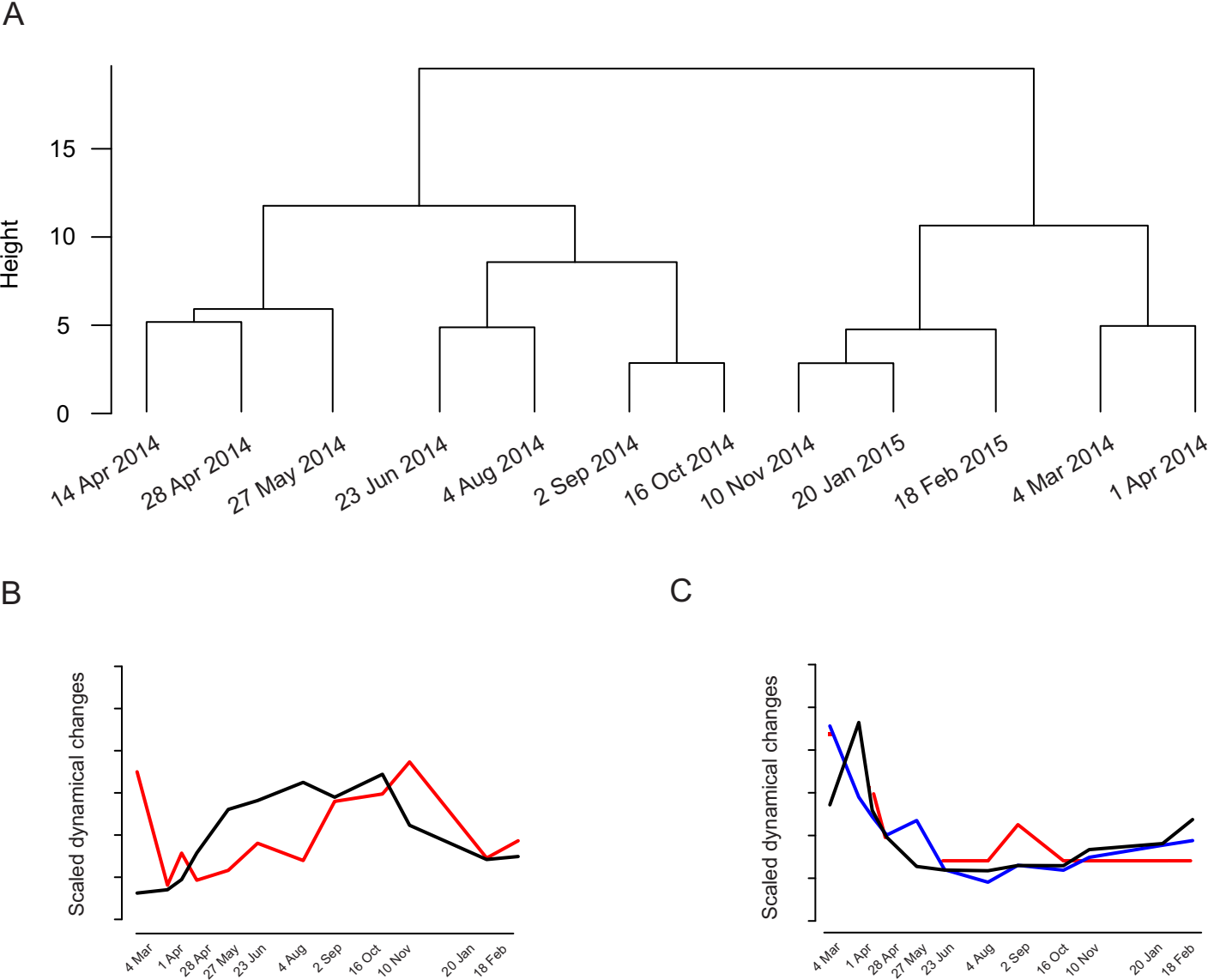


D




a
Protein-based cluster 2
63%
100%
b
Protein based cluster 20
65%
100%
c
22%
100%
d
DNA-VC-200 Contig_2869
DNA-VC-200 Contig_1859
DNA-VC-183 Contig_453
DNA-VC-206 Contig_709
DNA-VC-160 Contig_2873
DNA-VC-198 Contig_3073
DNA-VC-202 Contig_3051
DNA-VC-182 Contig_2797
DNA-VC-195 Contig_3109
DNA-VC-170 Contig_3095
DNA-VC-157 Contig_837 (partial)
DNA-VC-722 Contig_3179
DNA-VC-722 Contig_2819
DNA-VC-722 Contig_3689 (partial)
DNA-VC-720 Contig_1827
DNA-VC-722 Contig_2819
DNA-VC-712 Contig_2713
Mar Apr Apr Apr May Jul Aug Sep Oct Nov Jan Feb


Number of pairs
35
30
25
20
15
10
5
−10
−5
0
5
Distance from the highest host abundance (in sampling time points)


A
Height
15
10
5
0
14 Apr 2014
28 Apr 2014
27 May 2014
23 Jun 2014
4 Aug 2014
2 Sep 2014
16 Oct 2014
10 Nov 2014
20 Jan 2015
18 Feb 2015
4 Mar 2014
1 Apr 2014
B
Scaled dynamical changes
4 Mar
1 Apr
28 Apr
27 May
23 Jun
4 Aug
2 Sep
16 Oct
10 Nov
20 Jan
18 Feb
C
Scaled dynamical changes
4 Mar
1 Apr
28 Apr
27 May
23 Jun
4 Aug
2 Sep
16 Oct
10 Nov
20 Jan
18 Feb