From Trees to Clouds:
PhageClouds for Fast Comparison of ∼640,000 Phage Genomic Sequences and Host-Centric Visualization Using Genomic Network Graphs

Guillermo Rangel-Pineros, PhD1,2,1 Andrew Millard, PhD3 Slawomir Michniewski, PhD4 David Scanlan, PhD,4 Kimmo Sirén, PhD1 Alejandro Reyes, PhD2 Bent Petersen, PhD5,6,ii Martha R.J. Clokie, PhD3 and Thomas Sicheritz-Pontén, PhD5,6

Abstract

Background: Fast and computationally efficient strategies are required to explore genomic relationships within an increasingly large and diverse phage sequence space. Here, we present PhageClouds, a novel approach using a graph database of phage genomic sequences and their intergenomic distances to explore the phage genomic sequence space.

Methods: A total of 640,000 phage genomic sequences were retrieved from a variety of databases and public virome assemblies. Intergenomic distances were calculated with dashing, an alignment-free method suitable for handling massive data sets. These data were used to build a Neo4j® graph database.

Results: PhageClouds supported the search of related phages among all complete phage genomes from GenBank for a single query phage in just 10 s. Moreover, PhageClouds expanded the number of closely related phage sequences detected for both finished and draft phage genomes, in comparison with searches exclusively targeting phage entries from GenBank.

Conclusions: PhageClouds is a novel resource that will facilitate the analysis of phage genomic sequences and the characterization of assembled phage genomes.

Keywords: phage genomics, comparative genomics, genomic graph database

Introduction

Phages are viruses that target bacteria and were first described independently by Frederick Twort and Felix D’Herelle in the early 20th century.1 Since then, phages have been regarded as a potential therapeutic approach to treat bacterial infections, particularly in recent years due to a higher incidence of antibiotic-resistant pathogens.2,3 This has encouraged the isolation and characterization of novel phages that target clinically relevant bacteria, leading to the accumulation of isolated phage genomes in reference databases (Fig. 1).

1Section for Evolutionary Genomics, The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
2Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de los Andes, Bogota, Colombia.
3Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom.
4Warwick Medical School of Life Sciences, University of Warwick, Coventry, United Kingdom.
5Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia.
6Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark.
1ORCID ID (https://orcid.org/0000-0003-3848-4330).
2ORCID ID (https://orcid.org/0000-0002-2472-8317).

© Guillermo Rangel-Pineros et al. 2021; Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons Attribution Noncommercial License [CC-BY-NC] (http://creativecommons.org/licenses/by-nc/4.0/) which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and the source are cited.
FIG. 1. The number of complete phage genomes deposited in GenBank across time. The introduction of NGS technologies in the early 2000s was followed by an exponential increase in the number of phage genomes deposited in GenBank. The stacked bar plot on the left indicates the proportion of phage genomes that target different bacterial genera. The top eight targeted bacterial genera account for half of the phage genomes currently available in GenBank.

In addition to their potential as therapeutic agents, various studies have demonstrated the impact that phages have on biogeochemical cycling, bacterial ecology and evolution, and even potential roles in human health.1–6 Such studies, along with increasingly cheaper and accessible high-throughput sequencing technologies, have prompted the discovery of a myriad of phage genomic sequences.7 These sequences have primarily been derived from the identification of putative prophages in bacterial genomes or the detection of phage genomic sequences in metagenomic data sets or viral-enriched samples.8–10 This explosion of phage genomic data has uncovered an astonishing spectrum of genetic diversity.11 However, as the volume of phage genomic data grows, the estimation of genomic relatedness by alignment-based methods becomes increasingly impractical as these are slow and require large amounts of computing memory. Several alternative approaches have recently been developed to circumvent these limitations, among which the use of the MinHash algorithm became a popular choice for a range of tools.12–14 For instance, Mash uses MinHash to calculate the distance between a pair of genomes by estimating the Jaccard index for their combined kmer content, and produces a result that strongly correlates with the average nucleotide identity (ANI).12

Mash’s optimal use of computational resources was leveraged for estimating the distances between all genome pairs among 2333 phages from NCBI’s RefSeq.15 The study reported that using kmer and sketch sizes (Mash’s main parameters) of 15 and 25,000 resulted in the correlation of Mash distances and ANI values for the analyzed genomes, up to a Mash distance of <0.33.15 This demonstrated that tools such as Mash can be used to quickly and reliably estimate close evolutionary relationships between phages in massive data sets, where alignment-based methods become computationally impractical.16

In addition to estimating intergenomic distances for large phage data sets, it is imperative to select appropriate visualization strategies that aid the identification of clusters of related phages. Phylogenetic trees are commonly used to illustrate evolutionary relationships between biological entities, and they have been extensively used for phages when the inferred relationships are based on single or small sets of genes.17–19

More recently, a phylogenetic tree of phages in the order Caudovirales was generated with a collection of prevalent single-copy genes, which proved to be highly consistent with the taxonomy defined by the International Committee on Taxonomy of Viruses (ICTV).20 However, their use for showing relationships between phages has been argued against, as they fail to depict intercluster connections that result from the horizontal exchange of gene modules.21–22 Instead, other studies have showcased the use of networks as an alternative for representing phage evolutionary relationships.23,24 In fact, the algorithm used by the phage taxonomic classification tool vConTACT2 is based on the construction and analysis of gene-sharing networks.25 Despite the progress made so far, there is a need for resources that efficiently place novel phage genomic sequences within the context of the currently known phage diversity. Here we present PhageClouds, a novel approach based on the creation of a graph database that stores phage genomic sequences from a range of databases and their intergenomic distances. PhageClouds has a user-friendly interface allowing the user to search a database of 635,850 phage genomic sequences quickly and efficiently, and retrieve clusters of related phages (from hereon in referred to as phage clouds) based on a user-defined query.

Methods

Data set selection

The PhageClouds’ graph database was built using phage genomic sequences obtained from the following sources: NCBI’s GenBank database, the Gut Phage Database (GPD),10 phages from the Global Ocean Viromes (GOV) reported by the TARA Oceans consortium,9 phages in the Integrated Microbial Genome/Virus (IMG/VR) database7 from samples categorized as terrestrial, prophages from bacterial genomes in the Genome Taxonomy Database (GTDB),26 non-GOV entries in the Phages and Integrated Genomes Encapsidated Or Not (PIGEON) database27 the Cenote Human Virome Database (CHVD),28 The Gut Virome Database (GVD),29 phages from the cattle slurry virome,30 phages from horse feces viromes,31 and phages from marine samples taken at the U.K.’s south coast (MarineUKSouth)32 (Table 1).

Genomes of dsDNA phages were retrieved from NCBI’s GenBank database using the following Entrez query: db = “nucleotide”, term = “gbdiv PHG[prop]”. To filter out entries corresponding to incomplete genomes or genomic segments, the following terms were searched in each record
The collected phage genomic data and the calculated intergenomic distances were used to build the PhageClouds graph database. PhageClouds is an open-source native graph database that, unlike traditional relational databases, has a flexible structure that directly stores the relationships between the data. This type of database was selected as it allowed the modeling of our collection of phages and intergenomic distances as a network of nodes and relationships, hence enabling large-scale searches of phage clusters based on a predefined query.

Results and Discussion

Fast and structured illustration of host-specific phage sequence space

One of the main functionalities offered by PhageClouds is the ability to explore the phage genomic sequence space associated with a user-defined host. This approach will allow users to identify clusters of closely related phages that target a specific host, which provides an opportunity to rapidly explore and get a visual overview of the known phage space and diversity associated with that host.

To illustrate this, we searched for all phage clouds containing phages that target bacteria from the genus *Pseudomonas* and whose members are connected by intergenomic distances of no >0.15. The query used for searching these clouds identified all phage nodes whose metadata indicated they target *Pseudomonas* and all nodes connected to them with an intergenomic distance below the set threshold. Thus, the resulting clouds also included genomic sequences of phages either targeting other bacterial genera or lacking data on the host they target. After 7 min and 26 s, the search generated the graph illustrated in Figure 2, where nodes are proportionally sized to the genome size and colored according to the database where the corresponding phage was retrieved from.

### Table 1. List of Reference Databases Used for Building Our Graph Database

<table>
<thead>
<tr>
<th>Source</th>
<th>Sequence count</th>
<th>Date accessed</th>
<th>% targeted host</th>
<th>% country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenBank</td>
<td>17,062</td>
<td>July 29, 2021</td>
<td>87.5</td>
<td>62.1</td>
</tr>
<tr>
<td>GPD</td>
<td>142,809</td>
<td>January 1, 2021</td>
<td>23</td>
<td>92.7</td>
</tr>
<tr>
<td>TARA Oceans</td>
<td>195,728</td>
<td>November 2, 2020</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GTDB prophages</td>
<td>64,180</td>
<td>July 17, 2020</td>
<td>98.6</td>
<td>0</td>
</tr>
<tr>
<td>IMG/VR terrestrial</td>
<td>45,364</td>
<td>January 29, 2021</td>
<td>9.82</td>
<td>0</td>
</tr>
<tr>
<td>PIGEON</td>
<td>95,047</td>
<td>June 5, 2021</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GVD</td>
<td>15,330</td>
<td>June 23, 2020</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHVD</td>
<td>42,142</td>
<td>June 5, 2021</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Horse virome</td>
<td>1640</td>
<td>June 5, 2021</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MarineUKSouth</td>
<td>9915</td>
<td>June 5, 2021</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Slurry</td>
<td>6633</td>
<td>June 5, 2021</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Data on the number of sequences provided by each database, the last date they were accessed, and the percentage of entries that have information on targeted hosts and countries of isolation/detection are provided.

CHVD, Cenote Human Virome Database; GPD, Gut Phage Database; GTDB, Genome Taxonomy Database; GVD, Gut Virome Database; IMG/VR, Integrated Microbial Genome/Virus; PIGEON, Phages and Integrated Genomes Encapsidated Or Not.
The individual clouds displayed in Figure 2 may potentially be contrasting phage types that use different ecological strategies. For example, several jumbo phage clouds such as those containing the myoviruses phiKZ and PA5oct are easily identifiable, as well as other clouds that contain smaller phages such as the siphovirus, YuA, and the podovirus LUZ24. These clouds are not connected to each other at the selected intergenomic distance threshold, but many do contain well-characterized phages with very different infection strategies.

For example, YuA is an all-rounder phage that pursues a “leeching” strategy by actively depleting host metabolites rather than manipulating the host, whereas phiKZ tunes the host cell to establish the conditions it needs to replicate efficiently. By contrast, PA5oct hijacks the host's transcriptional environment by directing the cellular metabolism to suit its needs. While we do not know for sure, it may be that other related phages in the clouds interact in similar ways with their bacterial hosts to the better understood phages.

Most of the obtained clouds consisted largely of entries from GenBank corresponding to phages that target members of the genus Pseudomonas. However, there was an instance of a phage cloud in which the number of GenBank phages targeting other hosts was higher than the number of those targeting Pseudomonas. This cloud, highlighted with a red arrow in Figure 2, contained phages from the PIGEON database, many GenBank phages targeting bacteria from the genus Salmonella, and a single GenBank phage that in our graph database is annotated as a Pseudomonas phage (accession number MN871475).

At the time this article was prepared, this GenBank entry was described as a Pseudomonas phage, although it was labeled as an unverified entry. Thus, this might be a Salmonella phage instead, and we expect that the phages from the PIGEON database in this cloud also target this genus. This observation demonstrates how PhageClouds can be leveraged to identify incorrect annotations among entries in reference phage databases, which would help...
prevent the propagation of such errors when these reference databases are used in phage sequence characterization pipelines.

**Rapid search of phage clouds related to user-defined query phages**

Our graph database allows users to analyze a custom set of query phages to identify all phage clouds that they are associated with. To showcase the simplest scenario, we searched our graph database using the complete genome sequence of *Enterobacter* phage IME278 (GenBank accession number MW748991). The search was conducted using 20 cores and setting the distance threshold to 0.15 and 0.21. These values were selected because both favor primarily the selection of connections between phages from the same genus, while largely excluding connections between phages from different genera (Supplementary Fig. S1).

One search was performed by limiting the search space to reference phages from GenBank, and another search was conducted using all entries in our graph database. This focus on GenBank stems from our decision to regard this database as our gold standard for annotations, as its entries generally provide more metadata and are more carefully curated than entries in the other selected databases. We acknowledge that there can be errors in the annotations linked to entries from GenBank, but this should be minimal in comparison with the remaining databases. The running times for all these searches are indicated in Table 2, including the search of the complete graph database using an intergenomic distance threshold of 0.21 that took only 15 s.

A more complex scenario is the search of related sequences for a larger set of query phages. As an example, we searched the graph database using 79 phages from GenBank that had not been included among the records used to build our database. The search of phage clouds was first restricted to entries from GenBank and was carried out using intergenomic distance thresholds of 0.15 and 0.21. Table 2 indicates the time that these searches took, and the corresponding results are depicted in Figure 3A and B. At least one related phage from GenBank was identified for 56 out of the 79 query phages for an intergenomic distance threshold of 0.15, but this number increased to 63 when the search was conducted with a threshold of 0.21.

| **Table 2. Running Times for Searching Phage Clouds Related to Different Sets of Query Phages, Using Different Combinations of Search Parameters and 20 Computing Cores** |
|-----------------|---|---|---|---|
| 0.15<sup>a</sup> | 0.21<sup>a</sup> | 0.15<sup>b</sup> | 0.21<sup>b</sup> |
| GB<sup>c</sup> | GB<sup>c</sup> | WGD | WGD |
| Single-query phage | 00:00:10<sup>d</sup> | 00:00:10 | 00:00:11 | 00:00:15 |
| 79 new GenBank phages | 00:13:42 | 00:25:09 | 00:39:37 | 02:22:11 |

<sup>a</sup>Selected intergenomic distance threshold.
<sup>b</sup>Graph database entries included in the search.
<sup>c</sup>Time format h:min:s.
GB, GenBank; WGD, Whole graph database.

Therefore, these results indicate that limiting the search to entries from GenBank and a maximum intergenomic distance of 0.21 does not lead to the identification of at least one related phage for all 79 query phages. Perhaps extending the search space to all entries in our graph database would help achieve the goal of finding a closely related genomic sequence for all 79 query phages.

We repeated the search of phage clouds for the same set of 79 query phages but including all entries in the graph database and testing the same combinations of search parameters. Table 2 indicates the running times required for each search, and the corresponding results are depicted in Figure 3C and D. Using an intergenomic distance threshold of 0.15 resulted in the identification of closely related sequences for 10 of the 16 singletons from the GenBank-constrained search. Furthermore, increasing the threshold to 0.21 led to the identification of closely related sequences for all of the singletons.

Table 3 lists the best matches from the graph database that were identified for each of the 16 singletons. Our results demonstrated that, despite a significant increase in running time, searching the whole graph database led to the identification of closely related sequences for all the query phages.

Figure 3E illustrates the clouds retrieved for five of the singletons by the search conducted over the complete database with an intergenomic distance threshold of 0.21. As indicated in the figure, most of these singletons are annotated in GenBank as uncultured marine viruses (Fig. 3E). The figure revealed that all phage sequences connected to each of the singletons correspond to entries from databases other than GenBank. In fact, most of these sequences corresponded to entries from the TARA Oceans reference data set. This is expected as this data set consisted entirely of marine viral sequences.

Moreover, four of the singletons ended up connected to significantly smaller sequences from the TARA Oceans data set, and many of these were not connected to each other. Supplementary Figure S2 illustrates the alignment of all reference phage genomic sequences connected to the query phage (GenBank accession LC629609) in phage cloud 3 from Figure 3E, some of which were derived from the same sample (e.g., Station18_DCM and Station31_SUR). Thus, some of these reference sequences might correspond to different fragments of a phage genome that was not fully assembled when this data set was generated. This suggests that PhageClouds might be a useful resource for detecting genomic fragments that are derived from the same phage genome.

**Phage clouds reflect taxonomic groups of phages**

To demonstrate that the phage clouds reliably represent groups of closely related phages, we analyzed clouds that included phages of the family *Herelleviridae*. Until recently, members of this family were classified in the *Spounavirinae* subfamily within the family *Myoviridae*, but a series of complementary genomic/proteomic analyses demonstrated that the spounaviruses were markedly distinct from other members of the family *Myoviridae*. Therefore, these analyses supported the creation of the family *Herelleviridae* and the definition of its internal structure, meaning that subfamilies and genera within it have been defined based on genomic/proteomic relationships between member phages. Thus, we
considered that this family would be an ideal example to illustrate that phage clouds depict genuine relationships between closely related phages.

Figure 4 shows the clouds extracted with an intergenomic distance threshold of 0.15 and that contain at least one known member of the family *Herelleviridae*. The figure shows that all phages with known genus membership clustered together with phages from the same genus. The figure also highlights several instances of phages not classified in the family *Herelleviridae* that were tightly connected to clouds that represent different genera within this family.

The majority of these correspond to entries in NCBI currently labeled as “unclassified bacterial viruses,” such as the entries with accession numbers MW528836 and MN935200. However, the nodes highlighted with red arrows in Figure 4 are two examples of phages in our graph database that are classified in the family *Myoviridae*. While the record of the phage connected to the Schiekviruses was recently removed from GenBank (accession number LR760131), the phage connected to the Bequatroviruses is *Bacillus* phage BM5 that is currently classified in the family *Myoviridae* (accession number KT995479).

The taxonomic classification of *Bacillus* phage BM5 was not established by the ICTV. Thus, our results suggest that this phage represents an instance of a misannotated record within the GenBank, which might have resulted from the fact that members of the *Herelleviridae* were formerly classified as *Myoviridae* based on virion morphology. This observation is supported by results obtained using VIRIDIC, which demonstrated that the *Bacillus* phage BM5 is more closely related to members of the genus *Bequatrovirus* than to phages from all the genera currently classified within the family *Myoviridae* in NCBI (Supplementary Fig. S3).

Nevertheless, the ICTV established a minimum of 70% nucleotide identity over the full genome as the criterion for classifying phages in the same genus. As the largest percent identity observed between the *Bacillus* phage BM5 and a member of genus *Bequatrovirus* was 50.4% (Supplementary Fig. S3), our results do not support the inclusion of this phage within this genus. Nonetheless, the data in Supplementary Figure S3 demonstrated that *Bacillus* phage BM5 is more likely to be a member of the *Herelleviridae* rather than the *Myoviridae* family.

Figure 4 also illustrates many examples of phage genomic sequences from the other data sets in our graph database that connect to clouds representing a variety of genera within the family *Herelleviridae*. Among these are 31 entries from IMG/VR that are indeed currently classified as members of this family. The presence of the remaining phage genomic sequences in the retrieved clouds suggests that they could be members of the corresponding genera within the family *Herelleviridae*.

However, PhageClouds was not designed to be a tool for taxonomic classification of phage sequences. Based on the search of *Herelleviridae* clouds described here, it seems feasible that PhageClouds could help with the identification of phage clusters that correspond to taxa at the species or genus rank. Nevertheless, we would highly recommend the
use of additional tools designed to provide taxonomic annotations of phage genomic sequences at the mentioned taxonomic ranks, especially if their use is advised by the ICTV (e.g., VIRIDIC and vConTACT2).16

Rapid searching of assembled viromes

To demonstrate the capacity to rapidly compare a large number of phages against the database, 971 previously assembled viral operational taxonomic units (vOTUs) from the U.K. coastal waters (MarineUKSouth data set) were used as a query.32 The wall time for comparison of all 971 vOTUs against the complete graph database with an intergenomic distance threshold of 0.20 was 1 h, 41 min, and 11 s (40 cores used). Previously only 75 (7.7%) of the MarineUKSouth vOTUs had been identified to cluster with a database of *14,000 known phages using vConTACT2 (wall time *48 h with 32 cores).32 Here, we were able to associate 517 (53.2%) vOTUs from the MarineUKSouth data set with similar phages in other environmental databases (Supplementary Fig. S4).

It could be theoretically possible to run this analysis with vConTACT2, but the practical aspects of comparing ~640,000 phages using all-versus-all protein alignments prevent this from being run by an average user in a reasonable time frame. Instead, vConTACT2 could be used as a complementary analysis that attempts to identify more distant relationships between phage clouds, based on the genes shared between them. Furthermore, if a user is interested in analyzing a very large set of query phages (e.g., thousands of vOTUs), a sensible approach could entail the use of PhageClouds to quickly dereplicate and identify very close relatives from our graph database, thereby reducing the number of query phages to be analyzed with more sensitive alignment-based tools.

Of the 517 vOTUs that were associated with other phages, the closest relatives based on the calculated intergenomic distances were exclusively found in other marine viromes (Supplementary Fig. S4). Given that both viromes were constructed from seawater samples, this was not entirely unexpected. However, it was possible to quickly establish that very similar phages were found in geographically distant places. The vOTU NODE40_HQ was previously found to be present at station L4 and Plymouth Sound surface water. PhageClouds identified a cluster of similar contigs in two different stations (78 and 155) from TARA Oceans (Fig. 5A, B). Three of the four contigs from station 155 could be aligned to NODE40_HQ with an ANI of between 96% and 98%. The alignment against NODE40_HQ demonstrated that both phages had a conserved synteny in gene order and content (Fig. 5C). The use of PhageClouds helped to rapidly establish that a vOTU found in the water samples off Plymouth coast was very similar to viral contigs from TARA

<table>
<thead>
<tr>
<th>Query phage</th>
<th>Best match</th>
<th>Source database</th>
<th>Intergenomic distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC629455:uncultured_marine_virus</td>
<td>Station58_DCM_ALL_assembly_NODE_127_length_41467_cov_96.610693</td>
<td>TARA Oceans</td>
<td>0.108736</td>
</tr>
<tr>
<td>LC629456:uncultured_marine_virus</td>
<td>Station58_DCM_ALL_assembly_NODE_174_length_37974_cov_33.677971</td>
<td>TARA Oceans</td>
<td>0.101924</td>
</tr>
<tr>
<td>LC629458:Caudovirales_sp</td>
<td>Station56_SUR_ALL_assembly_NODE_61_length_88758_cov_21.840028</td>
<td>TARA Oceans</td>
<td>0.064562</td>
</tr>
<tr>
<td>LC629459:uncultured_marine_virus</td>
<td>Station18_DCM_ALL_assembly_NODE_544_length_22784_cov_15.229311</td>
<td>TARA Oceans</td>
<td>0.166993</td>
</tr>
<tr>
<td>LC629461:uncultured_marine_virus</td>
<td>Flavobacteriales_bacterium_NHFO01000063_phage12_64kb</td>
<td>GTDB prophages</td>
<td>0.128538</td>
</tr>
<tr>
<td>LC629467:uncultured_marine_virus</td>
<td>Bacterium_isolate_PAZJ01000031_phage54_57kb</td>
<td>GTDB prophages</td>
<td>0.077934</td>
</tr>
<tr>
<td>LC629474:uncultured_marine_virus</td>
<td>Station31_DCM_ALL_assembly_NODE_1914_length_19272_cov_27.587397</td>
<td>TARA Oceans</td>
<td>0.169876</td>
</tr>
<tr>
<td>LC629494:Caudovirales_sp</td>
<td>Station30_DCM_ALL_assembly_NODE_1322_length_15351_cov_25.741828</td>
<td>TARA Oceans</td>
<td>0.181003</td>
</tr>
<tr>
<td>LC629500:Siphoviridae_sp</td>
<td>Station58_DCM_ALL_assembly_NODE_568_length_22502_cov_143.650866</td>
<td>TARA Oceans</td>
<td>0.037512</td>
</tr>
<tr>
<td>LC629563:uncultured_marine_virus</td>
<td>Station100_SUR_ALL_assembly_NODE_1036_length_14465_cov_49.298959</td>
<td>TARA Oceans</td>
<td>0.123668</td>
</tr>
<tr>
<td>LC629575:uncultured_marine_virus</td>
<td>Station123_SUR_ALL_assembly_NODE_2449_length_10403_cov_69.936799</td>
<td>TARA Oceans</td>
<td>0.129898</td>
</tr>
<tr>
<td>LC629600:uncultured_marine_virus</td>
<td>Station76_DCM_ALL_assembly_NODE_3213_length_10752_cov_25.308030</td>
<td>TARA Oceans</td>
<td>0.108739</td>
</tr>
<tr>
<td>LC629609:Libanvirus_sp</td>
<td>Station18_DCM_ALL_assembly_NODE_18_length_107714_cov_7.165711</td>
<td>TARA Oceans</td>
<td>0.058169</td>
</tr>
<tr>
<td>LC629612:uncultured_marine_virus</td>
<td>PIGEON_EarthsVirome_17958</td>
<td>PIGEON</td>
<td>0.160845</td>
</tr>
<tr>
<td>MTO25940:Enquatrovirus_sp</td>
<td>Station168_1ZZ_ALL_assembly_NODE_477_length_89554_cov_52.179924</td>
<td>TARA Oceans</td>
<td>0.208478</td>
</tr>
<tr>
<td>MW822601:Synechococcus_phage_S-SRP02</td>
<td>IMGVR_UViG_3300035703_000158</td>
<td>IMG/VR (Terrestrial)</td>
<td>0.204246</td>
</tr>
</tbody>
</table>
FIG. 4. Herelleviridae clouds colored by genus. Clouds containing at least one member of the family Herelleviridae were retrieved from the graph database, using an intergenomic distance threshold of 0.15. Nodes are colored based on genus memberships, according to the annotations in the corresponding GenBank files. Gray nodes correspond to Herelleviridae phages without genus affiliation. Black nodes correspond to GenBank phages classified in other families or not currently classified at the family level. White nodes refer to phages from any of the other databases used to create the graph database. The red arrows highlight examples discussed in the text of GenBank phages classified in other phage families.

FIG. 5. Searching phage clouds for a set of 971 marine vOTUs. The search was conducted over the complete graph database using an intergenomic distance threshold of 0.20. (A) Illustrates a section of the obtained phage clouds (the complete set is illustrated in Supplementary Figure S3). Yellow nodes represent vOTUs from the U.K. coastal water viromes, and blue nodes represent entries from the TARA Oceans data set included in the PhageClouds’ graph database. (B) Shows a detailed view of the cloud highlighted in red from (A). The green and blue arrows point to the phage sequences present in the sequence alignment depicted in (C). vOTUs, viral operational taxonomic units.
station 155 (lat 54.5742, long -16.8345) that is ~1000 km apart. Furthermore, the station 155 contigs are likely fragments from a larger more complete phage genome.

Conclusion

PhageClouds is a computational resource that exploits the power and versatility of graph databases to offer a way of exploring the phage genomic diversity encompassed by published viromes. This tool will enable scientists to analyze their complete and draft phage genomes easily and efficiently by comparing them to a massive data set of ~640,000 reference phage sequences. In addition, PhageClouds is suitable for exploring phage diversity under a host-centric perspective, facilitating the identification of different groups of phages with different infection strategies.

Furthermore, inspection of phage clouds searched for a specific host simplifies the identification of potentially incorrect host annotations, which are easily propagated by the association with closely related phage genomic sequences. PhageClouds is also suitable for identifying genomic fragments from viromes that might be part of the same phage genome and ultimately supports the recovery of more complete genomes.

Currently, PhageClouds’ underlying algorithm and assumptions encompass equally all phage entries in our graph database, regardless of the type of nucleic acid in their genomes [i.e., dsDNA, ssDNA, ssRNA(+), ssRNA(−), dsRNA]. We acknowledge that phages with different types of nucleic acids might be associated with very different mutation rates, which suggests that the intergenomic distance thresholds used for finding clusters of related phages must differ between varying nucleic acid types.

Up to the time this article was prepared, and probably for the foreseeable future, most of the genomic sequences in public databases correspond to dsDNA phages, as evidenced by GenBank and IMG/VR, in which this type of phages correspond to ~97% of their entries.7 With this in mind and considering that one of PhageClouds’ main goals is to provide a resource for the quick and efficient comparison of a massive set of reference phage genomic sequences, we thought that the current functionality of our tool would constitute a good starting point. Nonetheless, future versions of PhageClouds could look deeper into defining bespoke intergenomic distance thresholds for groups of phages with different types of nucleic acids.

PhageClouds is hosted on our online server and is accessible from any web browser, and thus, users of this resource do not require any experience running software from the command line. PhageClouds is part of our online infrastructure at https://phagecompass.dk and https://phageclouds.dk. The tool accepts phage genomic sequences in FASTA format as input to query the graph database and allows users to examine precalculated phage clouds filtered by a specific host.

Authors’ Contributions

T.S.P. and M.R.J.C. conceived the idea of PhageClouds, and G.R.-P. generated and analyzed the data. A.M., A.R., and K.S. contributed to the design and content of the graph database. B.P. established the PhageClouds’ computing and testing infrastructure. S.M., D.S., and A.M. generated the marine virome assemblies used for testing PhageClouds. All authors reviewed, proofread, and approved the final article. They hereby confirm that all the coauthors have reviewed and approved the submission of this article. They also confirm that this article has been submitted solely to this journal and is not published, in press, or submitted elsewhere.

Acknowledgment

All calculations were performed on the Danish National Life Science Supercomputer Computerome.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

The work done by G.R.-P. was funded by the Foundation Idella project “Advancing the Medical Future Through Unlocking the Complex Genetic Material of the Amazon Rainforest.” The work done by K.S. was funded by the Mælkeafgiftsfonden project “Metacheese.”

Supplementary Material

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4

References


Address correspondence to:
Thomas Sicheritz-Pontén, PhD
Center for Evolutionary Hologenomics
GLOBE Institute
University of Copenhagen
Øster Farimagsgade 5, Bygning 7, Room 7.1.20a
Copenhagen 1353
Denmark

E-mail: thomassp@sund.ku.dk