IMG/VR: A database of cultured and uncultured DNA Viruses and Retroviruses

Summary

Viruses represent the most abundant life forms on the planet. Recent experimental and computational improvements have led to a dramatic increase in the number of viral genome sequences identified primarily from metagenomic samples. As a result of the expanding catalog of metagenomic viral sequences, there exists a need for a comprehensive computational platform integrating all these sequences with associated metadata and analytical tools. Here we present IMG/VR (https://img.jgi.doe.gov/vr/), the largest publicly available database of isolate reference DNA viruses with over 260,000 computationally identified viral contigs from more than 6,000 ecologically diverse metagenomic samples. Approximately half of the viral contigs are grouped into genetically distinct quasi-species clusters, enabling prediction of microbial host(s) for 20,000 viral sequences, revealing 14 microbial phyla previously unreported to be infected by viruses. Viral sequences can be queried using a variety of associated metadata, including habitat type and geographic location of the samples, or taxonomic classification according to hallmark viral genes. IMG/VR has a user-friendly interface that allows users to interrogate all integrated data and interact by comparing with external sequences, thus serving as an essential resource in the viral genomics community.

The resource

The Integrated Microbial Genomes with Virus (IMG/VR) database is an integrated viral analysis system within the IMG with Microbiome samples (IMG/M) data management system.

IMG/VR provides the largest integration of viral sequences with associated metadata and allows users to explore these data to decipher biogeographical and habitat distribution patterns of viral species as well as traveling across all the identified hosts putatively infected with viral sequences. In addition, users can compare and analyze their sequences against IMG/VR’s data (including viral protein family models, viral cluster and singleton information, distribution patterns of similar viral sequences across the globe, percent of known and unknown genes per sequence, and information regarding viral taxonomy and putative viral-host(s)), integrated with a variety of analytical tools.

We anticipate that IMG/VR will become a reference resource for sequence analysis of viral genomes and viral contigs derived from metagenomic samples.
How to use IMG/VR

Browsing iVGs and mVCs via Viral Datasets

The search functionality in IMG/VR is similar to that in the IMG/M system. All isolate viral genomes (iVGs) can be accessed via “Quick Genome Search” (by typing the virus name or taxon identifier “Taxon OID”) or “Find Genomes” tab (selecting viruses in “Genome Browser” or “Genome Search” tools) (Fig. 1).

The predicted mVCs are stored as metagenome scaffolds and they remain under their corresponding metagenome datasets (i.e. metagenome “Taxon OID”). Thus, metagenome “Taxon OIDs” can also be accessed the same way that any iVG and specific mVCs can be retrieved from the “Scaffold Search” tool of the “Find Genomes” tab (Fig. 1).

![Figure 1. General IMG/VR search functionality](image)

**Figure 1. General IMG/VR search functionality.** Basic search tools from IMG/M’s top menu bar (dashed red boxes) can be used to access the viral content of IMG/VR. “Quick Genome Search” at the top menu can be used to query specific viral names, taxon identifiers or keywords. Alternatively, all isolate viral content can be retrieved from the “Find Genome” tab, either selecting *Viruses* (boxed in grey with yellow background) from the “Genome browser” display (bottom left panel) or searching for *Virus* (boxed in grey with yellow background) in the *Domain* filter of the “Genome Search” tool (bottom central panel). To search for metagenomic viral contigs users need to access first the metagenomic sample (using any of the above tools). Then the “Scaffold Search” tool can be used to select specific scaffolds (bottom right panel).

In order to further facilitate the identification and selection of viral sequences in IMG/VR, all iVGs and mVCs can be accessed from the left panel table (IMG Viral Content) available from the entry page (*Home* tab) (Fig. 2a). This entry point enables browsing all viral
datasets in the context of their associated samples and corresponding metadata, e.g. habitat type or depth of the metagenome sample from which a viral sequence was identified (Fig. 2b). This table provides information about the total number of viral contigs per sample in IMG, allowing a quick identification of the samples with the largest number of viruses. Similar to other tables in IMG, the results can be exported in a tab-delimited text format compatible with other tools for metagenomics analysis, as well as R and Microsoft Excel (Fig. 2b). By clicking on the “Viral Contig Count” number from the previous table, users can examine the list of viral contigs from individual samples (Fig. 2c). The information displayed for a selected contig or group of contigs includes: scaffold identifier (Scaffold ID), gene count per contig (Gene Count), contig length (Sequence Length bp), guanine and cytosine content (GC Content), percent of genes per contig covered with viral protein families (Perc VPFs), viral species name identifier (Viral Cluster; detailed in “Sequence grouping” section and Supplementary data), predicted host and method of prediction (Host Detection; detailed in “Host-virus identification” section), taxonomic assignment at different levels based on POGs (Supplementary data), and the putative retrovirus sequences (Supplementary data).

Figure 2. Browsing IMG/VR viral datasets. (a) Total counts and access to the list of viral sequences from isolate viruses, metagenomic viral contigs, or their combination (dashed red oval). (b) Detailed table from “Total Viral Datasets” link displaying study and sample name, Taxon OID, habitat information, and number of metagenomic viral contigs. (c) List of viral metagenomic contigs found in a single sample. Columns in (b) and (c) can be sorted by clicking on the column header, and different filters can be used for specific searches (black boxes). Tables can be also exported in a tabular format by using the Export button (grey box with yellow background).
Browsing mVCs via environmental metadata

Metagenomic viral contigs can be viewed in relation to different environmental metadata associated with each sample. Two distinct curated environmental classifications systems are displayed at the bottom of the IMG/VR landing page, the ecosystem and the habitat type classification (Fig. 3).

The ecosystem classification is based on a previously developed five-tier hierarchical classification system (18). All metagenome data sets are organized in three main classes of the top ecosystem tier: *engineered*, *environmental*, and *host-associated*; and then further divided into sub-tiers called ecosystem category, ecosystem type, ecosystem subtype, and specific ecosystem (Fig. 3a). Currently, 78.3% of the mVCs belong to environmental samples, while 16.3% and 5.4% correspond to host-associated and engineered, respectively. Users can navigate through all samples at once or just reduce the search to any specific ecosystem class or category (i.e. “Environmental Terrestrial” Fig. 3b), and from there, select particular types, subtypes or specific ecosystems. The habitat type classification is based on 11 distinct manually curated habitat terms (e.g. air, freshwater, marine, host-associated human, host-associated plants, terrestrial soil) previously described. This classification allows the selection of mVCs from samples that belong to a single habitat type (Fig. 3c).
Figure 3. Accessing metagenomic viral contigs via associated environmental metadata. (a) Distribution of metagenomic viral contigs per ecosystem and ecosystem category information of samples according to GOLD classification. When a category is selected (e.g. Terrestrial samples -boxed in dashed red) a new table is displayed. (b) Detailed information of the selected Terrestrial samples. The total number of metagenomic viral contigs per sample (boxed in dashed red) can be viewed. Columns can be sorted by clicking on the column header, and different filters can be used for specific searches (black box). The table can be also exported in a tab-delimited text format by using the Export button (greyen box with yellow background). (c) Number of mVCs per Habitat Type category of the sample where the mVCs were found.

**Browsing mVCs via geographic location or human body site metadata**

Viral contigs can be viewed based on the geographic coordinates of a corresponding sample. This functionality is available primarily for environmental metagenomes and allows the selection of samples with specific location via “Marker Clusterer for Google Maps”, a javascript API utility library that creates and manages per-zoom-level clusters for large amounts of markers in Google Maps. Ultimately, as a user zoom in the map, the list of viral contigs that belong to those a sample(s) can be retrieved by clicking on a map pin and selecting the count next to the metagenome of interest for that location (Fig. 4a).

Additionally, all viral contigs identified in samples from the human body can be displayed by clicking on the “Show Human Body Sites” button (Fig. 4b). This option allows access to viral contigs derived from samples of any of the 5 main human body sites (nose, mouth, skin, intestine, and vagina), together with general statistics of these viruses per body site (Fig. 4c). From the default Human Body Sites summary table users can select all mVCs from a particular sample site or only those with a putative host (Fig. 4d).
**Figure 4. Maps of samples containing viral contigs from environmental and human-associated metagenomes.** (a) World interactive Google Map with a geographic location of metagenomic samples. All samples can be selected together or only those from any of the three major ecosystems. Map pins (in red) represent location counts of viral contigs and may contain multiple samples. Map pins are grouped into clusters and clusters themselves into larger clusters (bold number with a coloured halo based on number of members within the cluster) according to the Google Map javascript API utility library. As you zoom into any of the cluster locations, the number on the cluster decreases, and you begin to see the individual markers on the map. Zooming out of the map consolidates the markers into clusters again. (b) Human Body image showing the five body sites with available samples. All the metagenomic viral contigs identified in each body site can be accessed from the circles in the image. (c) Table provides information about mVC clusters/singletons, number of samples, and viral contigs with a host. (d) List of human skin samples with viral contigs. Columns can be sorted by clicking on the column header. Different filters can be used for specific searches (black box). This table can be exported by using the Export button (grey box with yellow background). Human body image credit: NIH Medical Arts and Printing.

**Browsing viral clusters and viral singletons**

Viral clusters and singletons together represent the entire viral diversity within IMG/VR. A total of 39,701 viral clusters and 122,665 singletons are available from the left panel on IMG/VR’s entry page (Fig. 5a). Together, these represent 162,366 viral quasi-species identified numerically with the prefix “vc_” or “sg_” depending if they belong to a viral cluster or remain as a singleton.

By clicking on the viral cluster or singleton identifiers the users can obtain information about the number of members in the cluster (“Viral Contig Count”), the number of samples in which they were found (“Sample Count”), the number of independent projects these samples belong to (“Study Count”), the proposed host (when detected, “Host”), and the sample’s habitat (“Habitat Type”) (Fig. 5b).

By clicking on a single viral cluster, all members of the cluster are displayed with several related metadata, including the number of genes per viral contig, contig length, GC content, host assignment, and taxonomic information (Fig. 5c).
Figure 5. Viral diversity in IMG/VR. (a) Accession link with the number of viral clusters or singletons available in the system. (b) Detailed table from “Viral Clusters.” (boxed in dashed red from panel a) showing the number of metagenomic viral contigs per cluster. The number of distinct samples, and unique projects (“Study Count”) is shown, besides information regarding host and habitat. (c) Viral cluster details table with host (via microbial CRISPR-spacer sequence matches) and taxonomic (via hallmark genes) information when available. Columns in (b) and (c) can be sorted by clicking on the column header, and different filters can be used for specific searches (black boxes). Tables can be also exported in a tab-delimited text format by using the Export button (grey box with yellow background).

Browsing viral clusters based on their host assignment

The third section of the left panel in the IMG/VR entry page shows the number of viral contigs associated with a host. Three different categories of host-linked contigs are provided (Fig. 6a). First, the number of isolate viruses experimentally assigned to a host is reported. There are currently 3,929 such viruses, which when accessed, are listed together with their corresponding host (Fig. 6b).

Second, the metagenomic viral contigs that bear a proto-spacer sequence match to a spacer from a microbial isolate genome (allowing a direct association virus-host at the species level) are reported. There are currently 8,091 mVCs which can be listed in a table grouped with their associated hosts. As an example, there are 131 different viral species (representing a total of 388 mVCs) putatively infecting Streptococcus oralis (Fig. 6c).
Finally, the total number of metagenomic viral sequences that can be assigned to a host (at the lowest possible taxonomic level) by projecting the host-virus information onto a viral cluster, is also presented. There are 13,947 in this category, whereby in the majority of the cases the virus-host link is at genus or species level. The microbial genera infected with the highest number of viral contigs are *Streptococcus*, *Veillonella*, *Fusobacterium*, and *Prevotella* (Fig. 6d). In ~9% of all assignments, the host connection is at a higher taxonomy range (ranging from family to phylum).

All the information from all the tables can be independently accessed by clicking on their corresponding links or could be exported in an excel tab-delimited text format by using the “Export” button.

![Image](image_url)

**Figure 6. Viral data sets with host assignment.** (a) Number of isolate viruses or metagenomic viral contigs with a predicted host. (b) “Isolate viruses with host” table sorted by the hosts infected by the highest number of viral genomes. (c) Top microbial host species containing metagenomic viral contigs. (d) Microbial hosts (at different taxonomic levels) with the highest number of metagenomic viral contigs assigned. Columns in (b), (c), and (d) can be sorted by clicking on the column header and different filters can be used for specific searches (black boxes). Tables can also be exported in a tab-delimited text format by using the Export button (grey box with yellow background).

**Sequence search**

Users can compare their sequences against the sequence data integrated into IMG/VR. Specifically, the sequences of all the viral contigs and all the spacer sequences from the isolate genomes can be queried by using the “Viral/Spacer Blast” option at the bottom of the home page (Fig. 7a). Both queries can be selected from “Blast Database” and rely on nucleotide BLAST searches (23) with customizable e-value identity cutoffs (Fig. 7b).
Matches against the viral database generate a list of viral sequences with a significant alignment based on the selected thresholds. These subject sequences can be directly accessed or selected-to-be-added to the Scaffold Cart, where their associated metadata are also provided. Similarly, matches of external viral sequences against the spacer database generate a list of host(s) which contain a CRISPR-spacer sequence with a significant alignment based on the selected cutoffs. These putative host(s) can be further explored by clicking on the host identifier. This redirects the user to detailed information of the spacer: source taxon name, location of the spacer within the CRISPR array, and spacer sequence (Fig. 7c).

Figure 7. Viral searches against IMG/VR databases. (a) Location of blast tool in IMG/VR (dashed red box). (b) User interface to blast sequences. Exclusively nucleotide sequences can be queried currently in the system. Sequence(s) must be added into the blank area. Users can blast their sequence(s) against any of the 2 databases integrated into IMG/VR: “Viral Sequence” or “Viral Spacer” and customize the e-value cutoff. (c) Example of a blast output of an external partial viral sequence (Streptococcus phage 858) against the spacer database. When a sequence hit (purple box) is selected, a new panel (“Viral Spacer”) is displayed showing details of the sequence spacer and the putative corresponding microbial host.