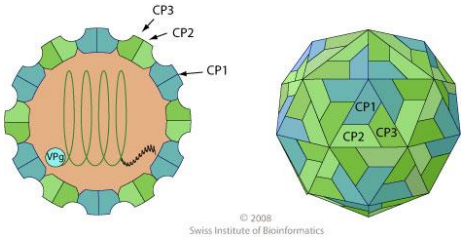

Virosaurus user manual

Version 2.0 (21.05.2019)

Virosaurus 11_18 files: https://drive.sib.swiss/index.php/s/RdxoiZZ7wp28pc6

Virosaurus (virus thesaurus) is a database for representative viral sequences of known genetic diversity, and curated in order to facilitate clinical metagenomics analysis.

Quickstart

The best way to use the database is have all reads mapped to Virosaurus sequences, then group sequences having the same **usual name**. Doing so, the results should be a small list of virus species, with extra data to assess quality of detection; like number of reads, genome coverage and sequencing depth.

Database contents

Virosaurus contains full-length genomes (monopartite genomes) or segments (segmented genomes) for all know vertebrate virus. Virosaurus covers all genetic diversity available in GenBank. All available sequences were clustered at 90% to remove redundancy in Virosaurus 90 (63,661 FASTAs, 102Mb); or clustered at 98% in Virosaurus98 (598,761 FASTAs, 232Mb). Many clusters can belong to the same virus species. For example, there are 25 Lassa clusters in Virosaurus90, 281 in Virosaurus98.

The FASTA header have been annotated with metadata to facilitate metagenomic analysis. For instance, viral nucleic acid is annotated as RNA, DNA or RNA/DNA, thereby improving interpretation from samples sequence based on either molecule.

In the Virosaurus release 11_18, herpesviridae and poxviridae sequences are split in genes rather than full genomes. This allows using incomplete genome sequences, and helps to mitigate the low number of complete genomes versus high variability for those families.

Licence :

FASTA format

Annotation are stored in FASTA header. The header contains 11 different topics annotated by a controlled vocabulary. Data comes from GenBank, ICTV, ViralZone and manual curation.

FASTA header:

> <GenbankID> :<SequenceID> ; **usual name=**<common clinical name>; **clinical level=**<SPECIES or GENUS>; **clinical typing=**<unknown or subtype name>; **species=**<species name>; **taxid=**<NCBI taxID>; **acronym=**<virus acronym>; **nucleic acid=**<DNA, RNA or DNA/RNA>; **circular=**<Y or N>; **segment=**<N/A or segment name>;

<GenbankID> Genbank accession number of the sequence displayed in the FASTA.

<SequenceID> By default Genbank accession number of the sequence displayed in the FASTA. If the displayed sequence is a portion of a GenBank entry, for example in the case of genes, the Sequence ID is a unique identifier like GENE_583-3988.

Usual name= Name of **clinical level** entity; If the scientific name is not commonly used, the common clinical name replaces species official name, for example parvovirus B19 is the usual name of *Primate erythroparvovirus 1* species. If **clinical level** =genus: genus name or acronym, for example all *Alphatorquevirus* usual name is TTV.

Clinical level= Gives the taxonomic level suggested to be relevant for usual clinical diagnostics. By default <species>, but can be at <genus> level like for TTVs or HPVs.

Clinical typing= Unknown by default. Otherwise contains data clinically relevant below species level. This can be genotypes (example:HCV) or qualifiers (polio enterovirus, High risk HPV, etc…). In rare cases of mixed cluster, several types are listed separated by a coma, this notably happens for some HPVs “low risk” and “undetermined risk” that are very similar.

Species= indicates the current official species name, as reported by International Committee on Taxonomy of viruses (ICTV): https://talk.ictvonline.org/taxonomy/ . In rare cases of mixed cluster, several species are listed separated by a coma, this notably happens for some segments of rotavirus A and C with are very similar within different species.

Taxid= Taxonomy identifier from NCBI taxonomy database: https://www.ncbi.nlm.nih.gov/taxonomy of the taxonomic entity at **species** level.

Acronym= Official acronym name of the **species,** as reported in ViralZone acronym list: https://viralzone.expasy.org/resources/Acronyms.xlsx

Nucleic acid= Nature of viral genome, either RNA or DNA for most viruses, RNA/DNA for retro-transcribing viruses (Ortevirales).

Circular= Y or N for yes or no. This is essential for to map efficiently reads at both extremities of the FASTA sequence.

Segment= N/A for monopartite viruses. For segmented genomes: official segment name as reported in ViralZone database: https://viralzone.expasy.org/

Creating a report using Virosaurus data

Virosaurus FASTA header is designed to simplify clinical metagenomics data report by gathering reads under each viral species with the optional addition of data like serotypes. The concept is to report reads associated to a <**usual name**> entity, rather than to individual sequences.

Virosaurus is clustered to lower the redundancy of sequences, which is rather high for HIV-1 and Influenza viruses. Each Virosaurus entry is a representative sequence from a cluster that can comprise between 1 to 20,543sequences.

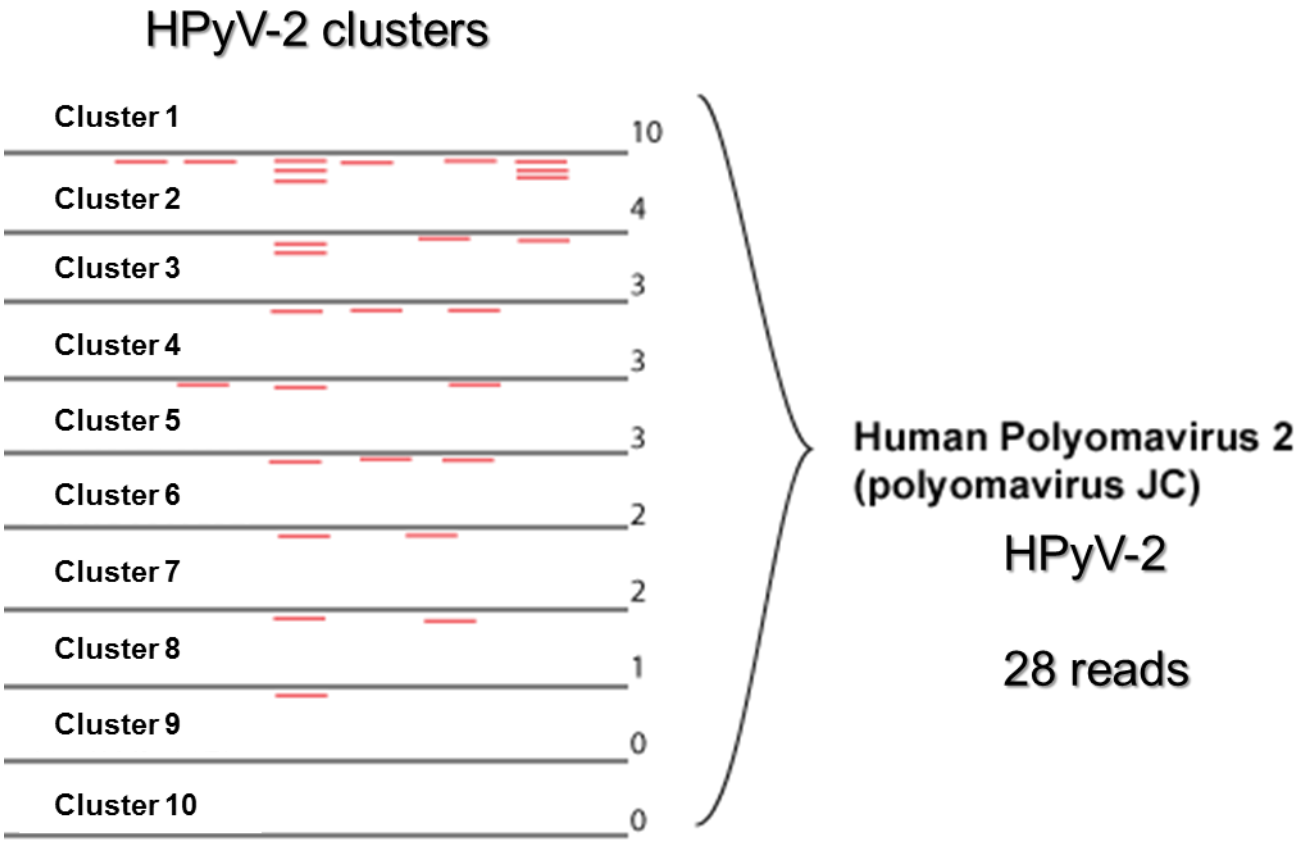

Figure 1: Example of reads grouped together under a species human polyomavirus 2. Here this virus genetic diversity is represented by 10 sequences in Virosaurus, representing 10 clusters of similar sequences. All clinical reads assigned to the “Human polyomavirus 2” <**usual_name**> FASTAs can ge added together, resulting in a total of 28 reads for HPyV-2. Doing so makes it easier to check the presence of viruses without having to look at a long list of similar viruses.

Virosaurus has been developed by a collaboration between SIB Swiss Institute of Bioinformatics (Vital-IT and Swiss-Prot groups), Université de Genève and Hôpitaux Universitaires de Genève.

Vital-IT
High Performance Computing Center