Virus Classification Pipeline (VCP; version 0.1)

Note This document describes how the Virus Classification Pipeline works.

1. Download virus sequences and the taxonomy database from GenBank (ftp://ftp.ncbi.nih.gov/genbank/)

```
$ perl viral_DB_prepare.pl -t download > download.sh
$ bash download.sh
```

These two commands will: 1) generate the download script (download.sh), and 2) download all virus sequences (GB_virus.gz) and the taxonomy database (names.dmp.gz and nodes.dmp.gz) from GenBank.

2. Run viral_DB_prepare.pl script to classify virus sequences into different kingdoms

```
$ perl viral_DB_prepare.pl -t category -c 1 GB_virus.gz 1>report.txt 2>&1
```

This will generate three files:
manual_hname_table.txt
manual_genus_table.txt
manual_desc_table.txt

3. Manual checking of the automatic classification

The automatic classification in step 2 may contain errors or unclassified viruses, which need to be manually processed.

3.1 manually correct host names associated with the virus sequences

The pipeline will classify some viruses according to their host names, if they cannot be classified using the genus information in ICTV (http://www.ictvonline.org/). However some host names in GenBank may not follow the standard description thus they cannot match entries in the taxonomy database. The file ***manual_hname_table.txt*** contains all the non-standard host names identified by the pipeline.

Example:

grape cultivar 6-23
grape cultivar 8612
grape cultivar 87-1

The standard host name for these entries should be 'Vitis vinifera' or 'wine grape' according to the GenBank taxonomy database.

The standard host name needs to be added after each of the non-standard names (separated by a tab key)
grape cultivar 6-23 Vitis vinifera

grape cultivar 8612 Vitis vinifera
grape cultivar 87-1 Vitis vinifera

3.2 Manually check the virus genus and classification

The automatic classification pipeline will generate some new genus classification information that is not presented in the ICTV website. For example, the becurtovirus (GI:169303562) is not present in ICTV and will be categorized as a plant virus according to its host (sugar beet). The pipeline will automatically generate a rule that becurtovius should be classified into the plant kingdom. The new genus classification information is stored in the file ***manual_genus_table.txt***. This file needs to be manually checked to ensure the accuracy.

3.3 Manually check the classification generated based on description and sequence similarity

After correcting host name and genus, some viruses still cannot be classified due to the missing of host features or having unrecognized genus names. However, most of them have nearly identical description, or show high sequence similarity to other classified viruses. Therefore these unclassified viruses can be classified by comparing their descriptions and sequences with classified viruses. For example, the sequence M18869 does not have host feature, and its genus name is "Small linear single stranded RNA satellites"; but it is described as "Cucumber mosaic virus satellite RNA". Another sequence X86421 with same description has been classified into plant virus. In addition, the sequence of M18869 is 100% covered by X86421 with 96% identity. Therefore, M18869 should be classified as a plant virus.

The pipeline automatically does this step and generates the file ***manual_desc_table.txt*** which stores the classification information of these viruses. However, this file needs to be manually checked to identify any potential wrong classifications by the pipeline.

The format of the file ***manual_desc_table.txt***:

1 - ID: GFLRNA3
2 - Description: Grapevine fanleaf virus satellite RNA (RNA3), complete cds.
3 - Same Description: Grapevine fanleaf virus
4 – Kingdom name by description
5 - No. of sequences with same description: 1
6 - Frequency of same description: 100.00
7 - ID of sequence with same description: GFLRNA1
8 - Kingdom name by blast
9 - best hit of blast
10- match length of blast
11- percentage identity
12- match score

Suggested method to manually check the file:

- A. Compare values in columns 4 and 8. Only need to check those with different values in these two columns.
- B. Check the blast result. Those with matching bases (column 10) less than 100 bp or with less than 90% identity (column 11) should be manually checked.

3.4 Add the manually corrected files to the classification

Next, the manually corrected files should be appended to previous files provided by the pipeline. The parameter **-v** indicate the version of GenBank (current version no. can be obtained at ftp://ftp.ncbi.nih.gov/genbank/GB_Release_Number).

```
$ perl viral_DB_prepare.pl -t patch -v 211
```

Three new update files will be generated
update_genus_table_v211.txt
update_hname_table_v211.txt
update_desc_table_v211.txt

4. Run viral_DB_prepare.pl script to classify viruses again

```
$ perl viral_DB_prepare.pl -t category -c 1 GB_virus.gz
```

5. Extract virus protein sequences

```
$ perl viral_DB_prepare.pl -t extProt GB_virus.gz
```

Two files will be generate.
vrl_genbank_prot: virus protein sequences
vrl_genbank_tab: virus nucleotide accession and protein accession

6. Remove redundancy in virus sequences (using plant viruses as an example)

```
$ perl viral_DB_prepare.pl -t unique –p 20 vrl_Plants_all.fasta -s 95
```

This will collapse redundant sequences with 95% sequence identity.

7. Retrieve proteins for each division (using vrl_Plants_u95 as an example)

```
$ perl viral_DB_prepare.pl -t genProt vrl_Plants_u95 vrl_genbank_prot
vrl_genbank_tab
```

The output protein sequences will be named as: ***vrl_Plants_u95_prot***