

Entrez database queries

Chris Stubben

April 22, 2010

Genome tables may also be created using two Entrez Utility functions. The `term2summary` function remotely queries the Genome Project database at NCBI using any valid combination of Entrez search terms and returns a genome table. Since detailed taxonomy information is not stored in the local tables, a typical search may include listing genomes projects by a taxonomy group like family, class or order. In addition, some fields in the genome tables like sequencing center may be incomplete and many other fields are missing. For example, this query returns a list of microbial genome projects which have sequence data in the Short Read Archive.

```
R> sra <- term2summary("genomeprj sra[Filter] AND Bacteria[ORGN]")
R> sra
```

A genomes data.frame with 651 rows and 4 columns

	pid	name	released
1	38739	Achromobacter xylosoxidans C54	<NA>
2	33685	Acidaminococcus fermentans DSM 20731	2010-01-15
3	34117	Acidaminococcus sp. D21	2009-03-10
4	29525	Acidimicrobium ferrooxidans DSM 10331	2009-04-30
5	37867	Acidovorax avenae subsp. avenae ATCC 19860	2009-12-08
...
651	29769	Yersinia ruckeri ATCC 29473	2009-06-01

	status
1	In Progress
2	Complete
3	Assembly
4	Complete
5	Assembly
...	...
651	Assembly

The `term2neighbor` function searches the Genome database and retrieves links to other genomes for a species (genome neighbors) in the Nucleotide database and then returns a

table listing accession numbers, deflines, released dates, and taxonomy ids. Viral genomes typically have one Reference sequence per species, and other strains are linked as Genome Neighbors. For example, Nipah virus is listed once in the virus table (NC_002728) and has 7 neighbors reported. To download those 7 neighbors, use the `term2neighbor` function shown in the next example. In addition, the function can also return the GenBank sequence that the reference was derived from using the `derived=TRUE` option.

Finally, if you are searching for a large group of viruses, it is often helpful to lookup the scientific name using the taxonomy ID in the table. The `taxid2names` function takes a taxonomy ID and returns the scientific name and lineage from the Taxonomy database. Using pattern matching, one can return the genus and plot released dates.

```
R> data(virus)
R> subset(virus, name %like% "Nipah*")

      name      released neighbors segments   refseq isolate  size
1428 Nipah virus 2000-06-01          7         1 NC_002728      - 18246
      proteins
1428          8
```

```
R> nipah <- term2neighbor("Nipah virus[orgn]")
R> nipah[, 1:2]
```

```
      acc                                          name
1 AJ564623 Nipah virus complete genome, isolate NV/MY/99/UM-0128
2 AJ627196 Nipah virus complete genome, isolate NV/MY/99/VRI-0626
3 AJ564622 Nipah virus complete genome, isolate NV/MY/99/VRI-1413
4 AJ564621 Nipah virus complete genome, isolate NV/MY/99/VRI-2794
5 AY988601          Nipah virus from Bangladesh, complete genome
6 AY029767          Nipah virus isolate UMMC1, complete genome
7 AY029768          Nipah virus isolate UMMC2, complete genome
```

```
R> buny <- term2neighbor("Bunyaviridae[ORGN]", derived = TRUE)
R> nrow(buny)
```

```
[1] 818
```

```
R> taxids <- unique(buny$taxid)
R> btax <- taxid2names(taxids)
R> genus <- gsub("(*Bunyaviridae; )(\w*)(.*)", "\\2", btax$lineage)
R> n <- match(buny$taxid, btax$taxid)
R> plotby(buny, genus[n], log = "y", lpty = "n", lcex = 0.7)
```

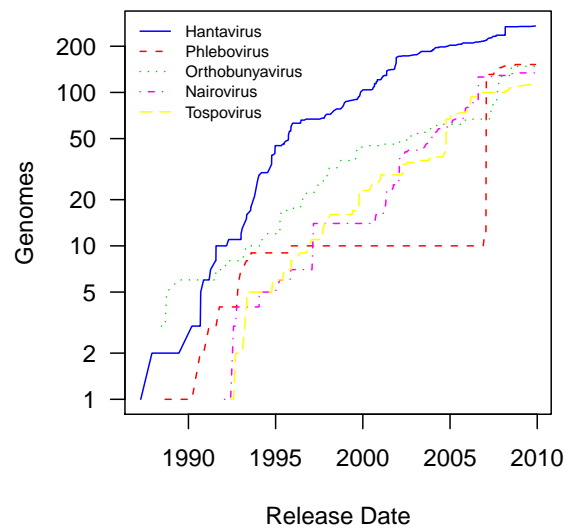


Figure 1: Accumulated number of genome sequences for vector-borne viruses in the family Bunyviridae.